UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE BIOLOGIA ANIMAL

Understanding ALS patients using Semantic Similarity

David Carriço Teixeira

Mestrado em Bioinformática e Biologia Computacional

Dissertação orientada por:
Professora Doutora Cátia Luísa Santana Calisto Pesquita
Professora Doutora Sara Alexandra Cordeiro Madeira

2019

Acknowledgements

First, I must thank professor Cátia Pesquita for her unwavering support throughout this work. Her knowledge, confidence, diligence, and patience provided me with the motivation needed for my best efforts. These were all qualities indispensable to my success, and were much appreciated. I also give my thanks to professor Sara Madeira for her advice and expertise, without which nothing here would have been possible, and for supplying me with all the data required for analysis and evaluation. I thank Manuel Figueiral and Sofia Pires, who helped me navigate and understand the data. I am obliged to the Fundação para a Ciência e a Tecnologia for their research grant, made available to me via funding from the LASIGE R&D Unit (UID/CEC/00408/2013) and by the SMILAX project (PTDC/EEI-ESS/4633/2014). Finally, I extend the greatest gratitude towards my family and parents, who selflessly stood by me along the best and worst of times, and for their ceaseless encouragement in these past, and so critical years.

Resumo Alargado

Nas últimas décadas, a pesquisa e a prática de biomedicina geraram e acumularam uma quantidade enorme e diversificada de conteúdo digital em literatura médica, incluindo artigos científicos, relatórios, notas médicas, e registros médicos eletrônicos (EMR) de dados clínicos que abrangem inteiros historiais de saúde de pacientes. Estas fontes integram uma coleção crescente de dados escritos em texto não estruturado, que precisam ser organizados e gerenciados para que possam ser analisados e interpretados adequadamente. Para este efeito, cada vez mais estão sendo implementadas técnicas computacionais e tecnologias da informação nas áreas da saúde. Em particular, a prospecção de dados permite a descoberta de padrões que podem ser usados para fazer previsões válidas. No entanto, as técnicas clássicas de prospecção têm dificulades em lidar com dados biomédicos não estruturados/semiestruturados, pois estes contêm um significado semântico profundamente enraizado, e que não pode ser detectado por meios de extração e análise direta de recursos.

O objetivo da web semântica e das tecnologias semânticas é tornar este tipo de informação mais explorável, fornecendo uma variedade de ferramentas que permitem uma tradução confiável por computador de conteúdo escrito, de forma a capturar o significado de palavras e frases. Este paradigma é suportado pelo armazenamento de dados semânticos em ontologias, que podem conter conhecimentos de vários domínios dentro e fora da assistência médica, e permitem a sua partilha através da anotação de termos nos dados. Por sua vez, ontologias, e anotações semânticas baseadas em ontologias, podem ser exploradas para várias tarefas de prospecção. Especificamente, e mais relevante para este trabalho, é o uso de medidas de semelhança semântica para calcular a semelhança semântica entre itens de dados. Assim como as técnicas de prospecção de dados geralmente dependem da proximidade ou distância entre os objetos encontrados nos dados, também podem analisar as relações semânticas entre as mesmas entidades, uma vez que elas são apoiadas pelo conhecimento da ontologia. Deste modo, se os dados puderem ser enriquecidos com o conhecimento externo de ontologias, uma prospecção mais informada poderá potencialmente retornar resultados mais precisos.

Este projeto abordou este desafio, desenvolvendo uma metodologia para analisar registros médicos de pacientes via integração com software e recursos semânticos, embora também possa ser generalizada para quaisquer entidades biomédicas que possam ser anotadas semanticamente com ontologias existentes. Um processo de agrupamento de semelhança semântica (SSC) foi concebido em torno de um pipeline de três etapas. Num primeiro passo, o pipeline cria uma rede semântica de ontologias que garante a cobertura ideal da anotação sobre os dados alvo, classificando as ontologias candidatas de acordo com o potencial da anotação e maximizando a quantidade de anotações que ela pode fornecer para os dados. A semelhança semântica é então calculada entre os pacientes como grupos de anotações, usando o SML como fonte de medidas de semelhança de entidades. O SML também oferece suporte para calcular semelhança semântica com múltiplas ontologias, por meio de um processo de reenraizamento que liga as ontologias a uma raiz virtual e as integra num um único grafo. O agrupamento foi realizado usando os algoritmos K-means ou Spectral Clustering do módulo Scikit-Learn do Python. Ambos são métodos populares com implementações sólidas, mas confiam em estratégias diferentes para encontrar estruturas compactas ou convexas ocultas nos dados, o que é útil quando não se tem conhecimento prévio sobre a forma do *cluster*. Valores de similaridades de pacientes são aceites na forma de uma matriz de semelhança, isto é, uma matriz de afinidade pré-computada e inserida como um núcleo para agrupamento. Além disso, foi desenvolvida uma ferramenta para elaborar uma descrição resumida do conteúdo semântico de um *cluster*, destacando seus elementos mais relevantes. Uma de-

scrição semântica de clusters terá como objetivo integrar ainda mais os dados de ontologia externa na análise do *cluster*, vinculando anotações de pacientes aos seus conceitos numa ontologia e estendendo-os com os seus ancestrais. Uma função de pontuação determina quais os conceitos que são mais significativos. Duas medidas de descrição de *clusters* foram definidas para este propósito - a primeira sendo uma função direta da presença do conceito dentro do *cluster* e da relevância semântica que possui na ontologia, e a segunda, baseada na análise de enriquecimento de conjuntos de genes (GSEA), atribui *P-Values* a cada conceito como uma medida de sobre-representatividade num *cluster*.

Estes métodos foram avaliados usando um conjunto de dados de 1376 pacientes com esclerose lateral amiotrófica (ELA), possuindo uma forte componente textual e uma heterogeneidade de sintomas generalizada entre os pacientes. Após os dados dos pacientes terem sido submetidos a um passo de pré-processamento, a metodologia para SSC foi aplicada. Os agrupamentos gerados foram então examinados para se entender o desempenho do SSC e como os pacientes de ELA podem ser classificados em grupos clínicos. Métodos de validação de *cluster* foram usados para medir a qualidade dos *clusters*, encontrar uma configuração ideal de *clusters* e avaliar os resultados finais. Métodos de validação extrínseca basearam-se numa *ground-truth*, ou seja, em *clusters* de referência, para comparar com os *clusters* de teste e atribuir-lhes uma pontuação que reflete a semelhança com esta referência. Por outro lado, métodos intrínsecos avaliaram os *clusters* relativamente às suas propriedades inerentes, examinando o quão separados e compactos eles são. Esta avaliação comparativa foi suportada por dois *clusters* de referência. Primeiro, usando uma abordagem padrão de agrupamento não-semântico - *Standard Approach* (SA) - no mesmo conjunto de dados para emular uma análise típica que qualquer operador poderia fazer num conjunto de dados de pacientes. Para este feito, os dados foram convertidos, usando uma codificação *one-hot*, para representações binárias que podem ser reconhecidas e manipuladas por algoritmos de *cluster* como K-means. Segundo, uma *ground-truth* usando grupos de progressão clínica (PG), onde os pacientes são estratificados em grupos separados por computação das suas taxas de progressão de ELA, um atributo que mede a rapidez com que a ELA progredide num paciente.

A qualidade dos *clusters* de SSC foi limitada devido ao baixo número e à imprecisão das anotações. Isto foi possivel verificar dada a cobertura restrita de conceitos pelas ontologias selecionadas e pelo facto de vários conceitos presentes no conjunto de dados, mas não no questionário, não terem sido anotados. O facto de o SSC não reconhecer dados negativos também contribuiu para uma falta geral de informações para o agrupamento de pacientes. Porém, comparando com a *ground-truth*, foi demonstrado que o SSC foi capaz de identificar alguns dos conceitos mais importantes que descrevem os *clusters*. Tanto o obtido por SSC, como pela abordagem não-semântica, não se aproximou do resultado esperado, pelo que o desempenho da metodologia proposta necessita de estudos subsequentes. Por outro lado, a descrição semântica de *clusters* baseados na análise de enriquecimento foi consistentemente capaz de fornecer *insights* significativos sobre a estratificação dos pacientes. As descrições de *cluster* são particularmente úteis, pois podem ser aplicadas a qualquer *cluster* de pacientes, semântico ou não. A descrição dos *clusters* de referência evidenciou como os sintomas e as regiões de início podem sugerir a velocidade de progressão da ELA. A capacidade de caracterizar rapidamente grupos de pacientes representa uma das maiores contribuições deste trabalho para a medicina personalizada, que depende de diagnósticos cuidadosos e detalhados de cada paciente.

Resumo

As técnicas clássicas de prospecção de dados têm dificulades a lidar com dados biomédicos não estruturados/semiestruturados, pois estes contêm um significado semântico profundamente enraizado em palavras e frases que não é detectado através da extração e análise diretas de recursos. Uma maneira de formalmente contextualizar dados é anotá-los com ontologias biomédicas e usar semelhança semântica sobre essas anotações para encontrar relações ocultas entre instâncias de dados. Deste modo, se os dados puderem ser enriquecidos com conhecimento externo, uma prospecção mais informada poderá, em princípio, retornar resultados mais precisos.

Este projeto abordou este desafio desenvolvendo uma metodologia para analisar registros médicos de pacientes por meio da integração com recursos e software semânticos. Uma *pipeline* de três etapas cria uma rede semântica de ontologias que garante cobertura semântica sobre os dados alvo, calcula a semelhança semântica entre pacientes com a aplicação SML (Semantic Measures Library), e agrupa pacientes usando algoritmos de clustering do módulo Scikit-Learn do Python. Além disso, foi desenvolvida uma ferramenta para elaborar uma descrição resumida do conteúdo semântico de um agrupamento, destacando os seus elementos mais relevantes.

Estes métodos foram avaliados usando um conjunto de dados de 1376 pacientes com esclerose lateral amiotrófica (ELA), possuindo uma forte componente textual e uma ampla heterogeneidade de sintomas entre pacientes. Os grupos de pacientes obtidos foram comparados, juntamente com uma baseline não-semântica, com grupos *ground-truth* de pacientes derivados das suas taxas de progressão de ELA. Foi demonstrado que a eficácia da metodologia proposta era fortemente dependente do número e da qualidade das anotações, mas também que os dados disponíveis não eram suficientes para detectar grupos de progressão. Apesar disso, as descrições de agrupamentos foram aplicadas com êxito em todas as abordagens, e forneceram informações úteis que evidenciaram pontos em comum entre o conteúdo semântico dos agrupamentos teste e da *ground-truth*. Por fim, esta metodologia pode ser generalizada para quaisquer entidades biomédicas que podem ser anotadas semanticamente com ontologias existentes.

Palavras Chave: Ontologias Biomedicas, semelhança Semântica, Esclerose Amiotrófica Lateral, Agrupamento

Abstract

Classical data mining techniques struggle with unstructured/semi-structured biomedical data, since it contains a deep rooted semantic meaning in words and sentences that cannot be detected through direct feature extraction and analysis. One way to give formalized context to data is by annotating it with biomedical ontologies, and using semantic similarity over these annotations to find hidden relationships between data items. Hence, if data can be enriched with external knowledge, a more informed mining can return, in principle, more accurate results.

This project addressed this challenge by developing a methodology to mine patient medical records via integration with semantic software and resources. A three-step pipeline builds a semantic network of ontologies that ensures semantic coverage over the target data, computes semantic similarity between patients with the Semantic Measures Library (SML) application, and clusters patients using clustering algorithms from Python's Scikit-Learn module. Additionally, a tool was developed to elaborate a summary description of a cluster's semantic content while highlighting its most relevant elements.

These methods were evaluated using a survey-based dataset of 1376 patients of Amyotrophic Lateral Sclerosis (ALS), possessing of a strong textual component and a widespread heterogeneity of symptoms among patients. Obtained patient clusters were compared, alongside a non-semantic baseline, against ground-truth patient groups derived from their ALS progression rates. It was shown that the effectiveness of the proposed methodology was heavily dependent on the number and quality of annotations, but also that the available data was not enough to detect progression groups. Despite this, cluster descriptions were successfully applied in all approaches, and delivered useful insights evidencing commonalities between their semantic content. Ultimately, this methodology can be generalized to any biomedical entities that can be semantically annotated with existing ontologies.

Keywords: Biomedical Ontologies, Semantic Similarity, Amyotrophic Lateral Sclerosis, Clustering

Contents

List of Figures

List of Tables

Acronyms

ACA All Common Ancestors.

ALS Amyotrophic Lateral Sclerosis.

ALSFRS-R ALS Functional Rating Scale Revised.

API Application Program Interface.

BMA Best Match Average.

BPD Borderline Personality Disorder.

CM Component Match.

CSV Comma-Separated Values.

DEG Differently Expressed Genes.

DO Disease Ontology.

ECAS Edinburgh Cognitive and Behavioural ALS Screen.

EMR Electronic Medical Record.

ESSO Epilepsy Syndrome Seizure Ontology.

FM Full Match.

FMI Fowlkes–Mallows Index.

FTD Frontotemporal Dementia.

GAF GO Annotation File.

GO Gene Ontology.

GSEA Gene Set Enrichment Analysis.

HPO Human Phenotype Ontology.

IC Information Content.

IDE Integrated Development Environment.

IOBC Interlinking Ontology for Biological Concepts.

KDD Knowledge Discovery from Data.

LASIGE Large-Scale Informatics Systems Laboratory.

LMN Lower Motor Neuron.

MICA Most Informative Common Ancestor.

NCBO National Centre for Biomedical Ontology.

NCIT National Cancer Institute Thesaurus.

NIFSTD Neuroscience Information Framework Standard Ontology.

NLP Natural Language Processing.

OBO Open Biomedical Ontologies.

ONTOAD Bilingual Ontology of Alzheimer's Disease and Related Diseases.

OWL Web Ontology Language.

PAM Partitioning Around Medoids.

PG Progression Groups.

PM Partial Match.

rDAG Rooted Directed Acyclic Graph.

RDF Resource Description Framework.

SA Standard Approach.

SML Semantic Measures Library.

SNOMEDCT Systematized Nomenclature of Medicine Clinical Terms.

SSC Semantic Similarity Clustering.

TSV Tab-Separated Values.

UMLS Unified Medical Language System.

UMN Upper Motor Neuron.

URI Uniform Resource Identifier.

VSM Vector Space Models.

WSD Word Sense Disambiguation.

XML Extensible Markup Language.

1 Introduction

1.1 Motivation

In the last few decades, biomedical research and practice have generated and accumulated a huge and diverse amount of digital content from medical literature, from scientific papers, reports, or physician notes, to electronic medical records (EMR) of clinical data spanning an entire person's health history [1, 2]. These sources integrate a large and growing collection of unstructured text data, which needs to be organized, curated, and managed so that it can be properly analysed and interpreted. However, the workload of these tasks is above the efforts of human labour.

Computational techniques and information technologies are being increasingly implemented in healthcare for this exact purpose. Data mining in particular, enables the discovery of patterns and relationships in data (i.e., knowledge) that can be used to make valid predictions [3]. In turn, by diagnosing and predicting diseases, the health service provided to patients can improve, while saving time and expense on incorrect treatments. Another important application of healthcare mining is in choosing the best treatment option for a specific patient, by comparing among all possible treatment techniques [4]. Personalized medicine itself is an emerging subject, a medical approach in which patients are stratified based on their disease description, so as to base medical decisions on individual patient characteristics. It has enabled early disease diagnosis, prevention, and management, resulting in a reduction of healthcare costs [5].

Data mining is grounded in other disciplines that play an important role in the same regard. For example, machine learning is a method of artificial intelligence that can learn relationships from the data without the need to define them a priori. It is especially useful when there is lack of formal codes, or there is poor definition of knowledge, and has been used for the prognosis of diseases, prediction of disease progression, or extraction of medical knowledge [6]. More recently, deep learning has shown improvements over conventional machine learning, and has been successfully applied to predict diseases from patient EMR [6].

Nevertheless, there are several technical obstacles to the mining of medical data, related to its inherent properties and uniqueness. As stated, raw medical data is heterogeneous, meaning that it has a plurality of formats, and while these provide unprecedented opportunities for research, their integration in a single data base and for a common purpose is complicated and manually unfeasible. Further, physician notes are often written in unstructured free-text using different grammatical constructs to describe relations between medical entities or different names used for same disease i.e., written in natural language that is difficult to standardize and mine. Finally, diseases and medical concepts themselves don't have a canonical form, and many semantically distinct expressions can have the same, or similar, medical meaning. [4, 7]. Many of these challenges can be addressed by semantic web technologies.

The goal of the semantic web and semantic technologies is to make web-accessible information and services more exploitable, by providing a range of tools that enable a reliable computer translation of written content to capture the meaning of words, and identify patterns in multiple and complex data. Such a paradigm is supported through the storage of semantic data in ontologies - rich conceptual schemas - like the National Cancer Institute Thesaurus (NCIT) [8] or the Interlinking Ontology for Biological Concepts (IOBC), which contain knowledge from a variety of domains within and out of healthcare, and allow its sharing through annotation of dataset terms. Thus, data integration and annotation go hand-in-hand, and in particular for complex multimodal datasets, annotation with single ontologies is often not sufficient [9]. The use of multiple ontologies is becoming more frequent, including for example, in the domain of phenotype descriptions [9, 10]. Employing data mining techniques over semantically annotated data is still a challenge, since classical techniques rely on vector data and ontologies are modelled as graphs. To overcome this, semantic data needs to be represented in a more amenable way.

One option, since as data mining techniques often rely on the closeness, or distance between objects found in data [11], is to rely on the similarity between the same given entities once these are backed by ontology knowledge. The semantic similarity of data can then be quantified through a variety of similarity measures following a manifold of strategies. So far, semantic similarity has seen enough success to be a well established field, although there are still issues it strives to solve, such as the adaptation and optimization of measures towards datasets of ever increasing size [12], or dealing with entities annotated with multiple ontologies [13].

While there have certainly been plenty of examples of semantic similarity measures being used to cluster biomedical data (e.g., proteins and genes), its application in the context of medical and disease records is still left unaddressed. The present work aimed to join semantic similarity and data mining into a clustering application directed at patient records. Here it is first tested on patients with Amyotrophic Lateral Sclerosis (ALS).

ALS is a neurodegenerative disease targeting the upper and lower motor neurons, leading to progressive and diffuse paralysis and eventual respiratory death. The ALS Therapy Development Institute estimates that there are 450,000 people worldwide with ALS, and its incidence in Europe has been recorded between 2 or 3 for every 100,000 individuals [14]. The underlying causes of ALS are poorly understood, as it manifests with an heterogeneity of clinical and genetic symptoms and a highly fluctuating survival duration. These factors have made ALS presence and clinical outcome difficult to predict among patients, causing numerous failed clinical trials, and slowing curative discovery [15].

Finding new insights to stratify ALS patients may yet lead to a more effective prognosis prediction, and an improved patient-personalized treatment. With these goals in mind, a novel methodology was developed for multiple-ontology semantic similarity clustering (SSC) of patient data, and implementing a semantic description of patient clusters based on the ontology content they contain. The usefulness of this approach was tested against a large survey based dataset, covering hundreds of ALS patients and their disease features.

1.2 Goals

This work aimed to:

- Develop new clustering analysis techniques for unstructured and complex survey based biomedical data, relying on ontology annotations and semantic similarity;
- Apply and test these methods to a case study on ALS patient data.

1.3 Contributions

This work has made the following contributions:

- A method to effectively annotate large collections of patient data through their survey content;
- A novel methodology to cluster patients by measuring their semantic similarity;
- A novel semantic approach to cluster interpretation;
- The implementation and application of these methods on real-world ALS patient data.

1.4 Document Structure

This document is organized as follows:

- Chapter 1, Introduction, outlines the motivation, goals and contributions of this work.
- Chapter 2, State of the Art, provides an overview of recent, and other relevant works, shaping the current stage of scientific development in data mining, semantic similarity, and their combined use in clustering data.
- Chapter 3, Methodology, details the general architecture, requirements, steps, and implementation of a novel patient clustering and cluster description approach based on semantic similarity.
- Chapter 4, Evaluation, explains how the methodology was evaluated, and expounds the results of its application on real-world patient data.
- Chapter 5, Conclusions, presents the main conclusions, and sets possible guidelines for future works to follow.

2 State of the Art

Knowledge Discovery from Data (KDD) is the overall process that converts raw data into useful information through a series of steps [11], represented in figure 2.1:

1. **Data cleaning** - to remove noise and inconsistent data.
2. **Data integration** - where multiple data sources may be combined.
3. **Data selection** - where data relevant to the analysis task are retrieved from the database.
4. **Data transformation** - where data are transformed and consolidated into forms appropriate for mining by performing summary or aggregation operations.
5. **Data mining** - an essential process where intelligent methods are applied to extract data patterns.
6. **Pattern evaluation** - to identify the truly interesting patterns representing knowledge based on interestingness measures.
7. **Knowledge presentation** - where visualization and knowledge representation techniques are used to present mined knowledge to users.

This work focuses on adapting and applying existing techniques belonging to different steps of the process in an effort to handle the challenges of mining complex and heterogeneous biomedical data. The emergence of the Biomedical Semantic Web, with multiple ontologies and linked data resources represents an opportunity to address these challenges, and in this work, it is the foundation of our approach. In **Data Integration**, biomedical ontologies are used to annotate the data, integrating it with the state of the art biomedical knowledge encoded in ontologies. In **Data Mining**, semantic similarity based on ontologies is used to cluster the data. In **Pattern Evaluation and Knowledge Presentation**, semantic annotations of the underlying data are used to represent and highlight relevant features of the clusters.

This chapter is organized as follows. Section 2.1 introduces the concepts of ontologies and reviews techniques for the semantic annotation of data. Section 2.2 details semantic similarity based on single and multiple ontology data, addressing viable metrics for similarity, and introducing software that implements these into computational tools. Section 2.3 reviews some essential core concepts on data mining and the more relevant clustering techniques used in this work, along with useful machine learning tools. Section 2.4 reviews related work on semantic similarity-based clustering - the main focus of this work.

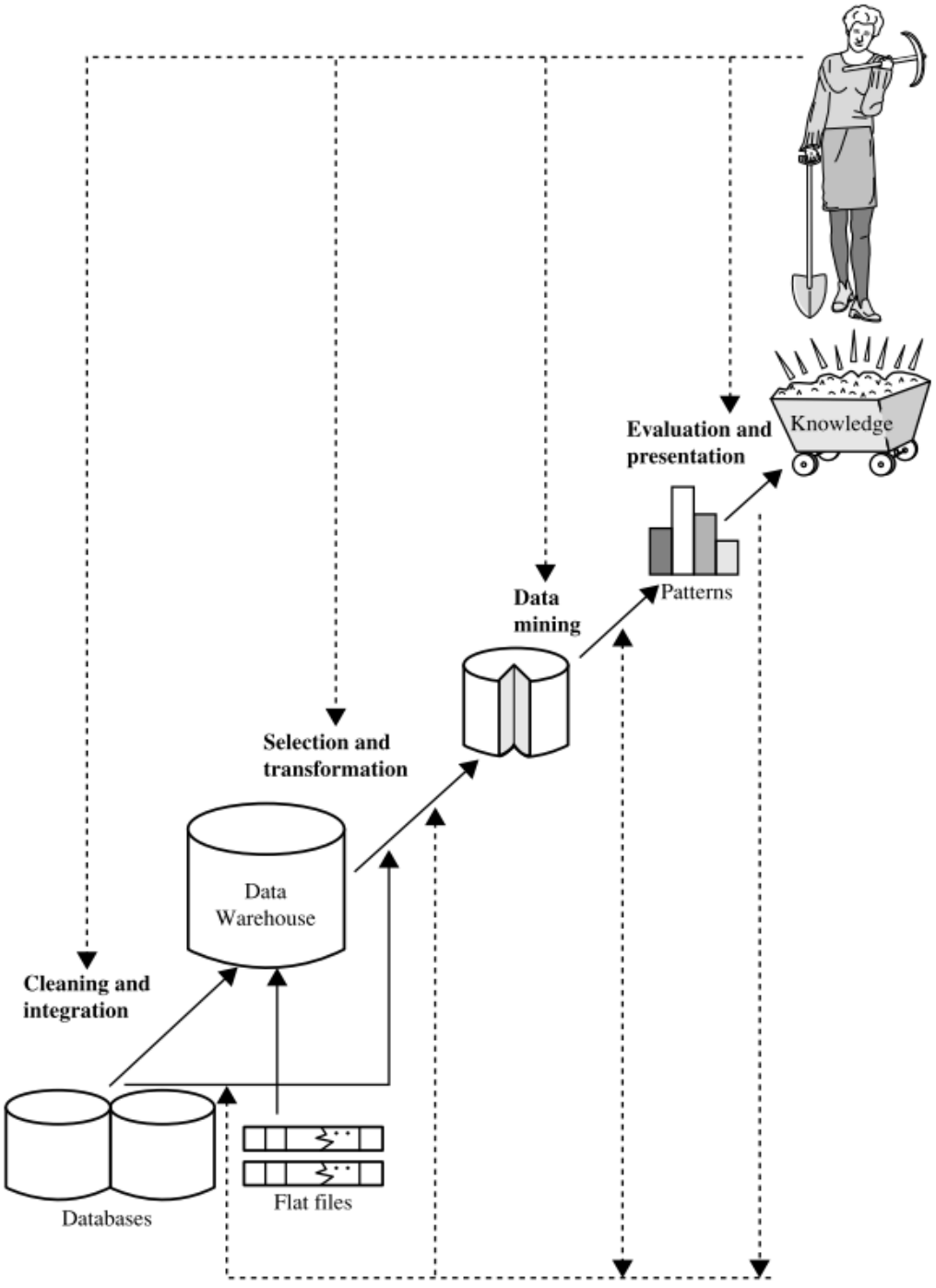

Figure 2.1: The process of Knowledge Discovery from Data [11].

2.1 Biomedical Semantic Web

The growing intricacy and size of biomedical research data has led to difficulties in information retrieval and analysis. To make the content of these documents more accessible to computational data-mining tools, the semantics, i.e. the meaning of a word, phrase, or text[1], embedded in them must first be translated into readily usable knowledge [16].

The semantic web was centered around the same problem, and sought to bring an additional, formal (machine-readable), structure to documents by enriching them with semantic annotations. A semantic annotation works much like a conventional annotation, i.e., a keyword assigned to a resource that, implicitly, describes a particular property, except that it is further extended to include semantic meta-data. It acts both as a label to terms found in data, and as a pointer to their formal definitions, carrying knowledge from latter to former [10, 17]. Semantic knowledge can be found and extracted from a knowledge base, usually in the form of an ontology.

2.1.1 Ontologies

Ontologies, in their original philosophical sense, are a branch of metaphysics aiming to study existence and the structure of the world, by developing hierarchical categorizations for the different kinds of entities that exist in it, and the features that distinguish them. Following similar guidelines, data science redefined ontology as an engineering artifact - a model of some aspect of the world, which contains and describes concepts belonging to it [10]. This adaptation has brought a new paradigm for ontologies.

Conceptually, an ontology is a semantic vocabulary covering a domain of variable scope and detail, and whose content reflects the consensus of specialists or communities dealing in the field in question. Domains are expressed mainly through classes referencing entities, and the relationships found between them (e.g. “part of”, “is a”) [9]. These in turn are represented in ontologies through axioms (example in Figure 2.2), or statements, which make their intended meaning explicit and generalized to the broadest possible situation, as as opposed to any particular state of affairs [18].

Axioms are thus instrumental in creating precise and unambiguous descriptions in an ontology. However, formalized ontologies are also able to encode these into a machine-processable form, using knowledge representation techniques [10]. This is one of the most useful features of ontologies. Information systems stand to benefit much from this property, since it essentially allows knowledge to be shared, processed, reused, captured and communicated [19], and enables computational applications to better understand ontology data by conducting reasoning tasks in order to make decisions [18].

[1] https://www.lexico.com/en/definition/semantics

Knowledge representation in ontologies was originally developed as a requirement for the semantic web, incorporating previous advances in artificial intelligence research. This resulted in the implementation of specialized ontology languages to express ontology components [18], the most prevalent being:

- **Resource Description Framework (RDF)** - Created as an early effort for the standardization of meta-data. It provides a way to link three uniform resource identifiers (URIs) to specify a pair of entities and a relationship between them (forming an RDF "triple") [20]. Though a valid ontology language, RDF is limited in expressiveness, i.e., by what kind of relationships it can convey [18].
- **Web Ontology Language (OWL)** - Developed in accordance with international World Wide Web Consortium (W3C) standards to be a more expressive language. Its basic components are individuals (the basic elements of the domain), classes (describe individuals having similar characteristics), and properties (describe relationships between pairs of individuals) [10]. Currently, OWL is the most prominent language used to express ontologies for their use in the Web [18].
- **Open Biomedical Ontologies (OBO) Flatfile** - Defines stanzas to describe each element in an ontology. A stanza is introduced by a name identifying the type of element being described, followed by lines containing tags and values, which describe relationships and other properties of elements. OBO uses a simple textual syntax that was designed to be compact, readable by humans, and easy to parse [21].

```
<owl:Class rdf:about="http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#C34373">
        <rdfs:subClassOf rdf:resource="http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#C4802"/>
        <NHC0>C34373</NHC0>
        <P106>Disease or Syndrome</P106>
        <P108>Amyotrophic Lateral Sclerosis</P108>
        <P207>C0002736</P207>
        <P366>Amyotrophic_Lateral_Sclerosis</P366>
        <P90>ALS</P90>
        <P90>Amyotrophic Lateral Sclerosis</P90>
        <P90>Lou Gehrig Disease</P90>
    </owl:Class>
    <owl:Axiom>
        <owl:annotatedSource rdf:resource="http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#C34373"/>
        <owl:annotatedProperty rdf:resource="http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#P90"/>
        <owl:annotatedTarget>ALS</owl:annotatedTarget>
        <P383>AB</P383>
        <P384>NCI</P384>
    </owl:Axiom>
    <owl:Axiom>
        <owl:annotatedSource rdf:resource="http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#C34373"/>
        <owl:annotatedProperty rdf:resource="http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#P90"/>
        <owl:annotatedTarget>Amyotrophic Lateral Sclerosis</owl:annotatedTarget>
        <P383>PT</P383>
        <P384>NCI</P384>
    </owl:Axiom>
```

Figure 2.2: Example of axioms from the NCIT ontology in OWL format.

Other than enabling human and machine comprehension of data, ontology axioms can give rise to a graph structure, which is visualized and thought of as a semantic network displaying ontology concepts as nodes, and their relationships as arcs [18] (see Figure 2.3 below).

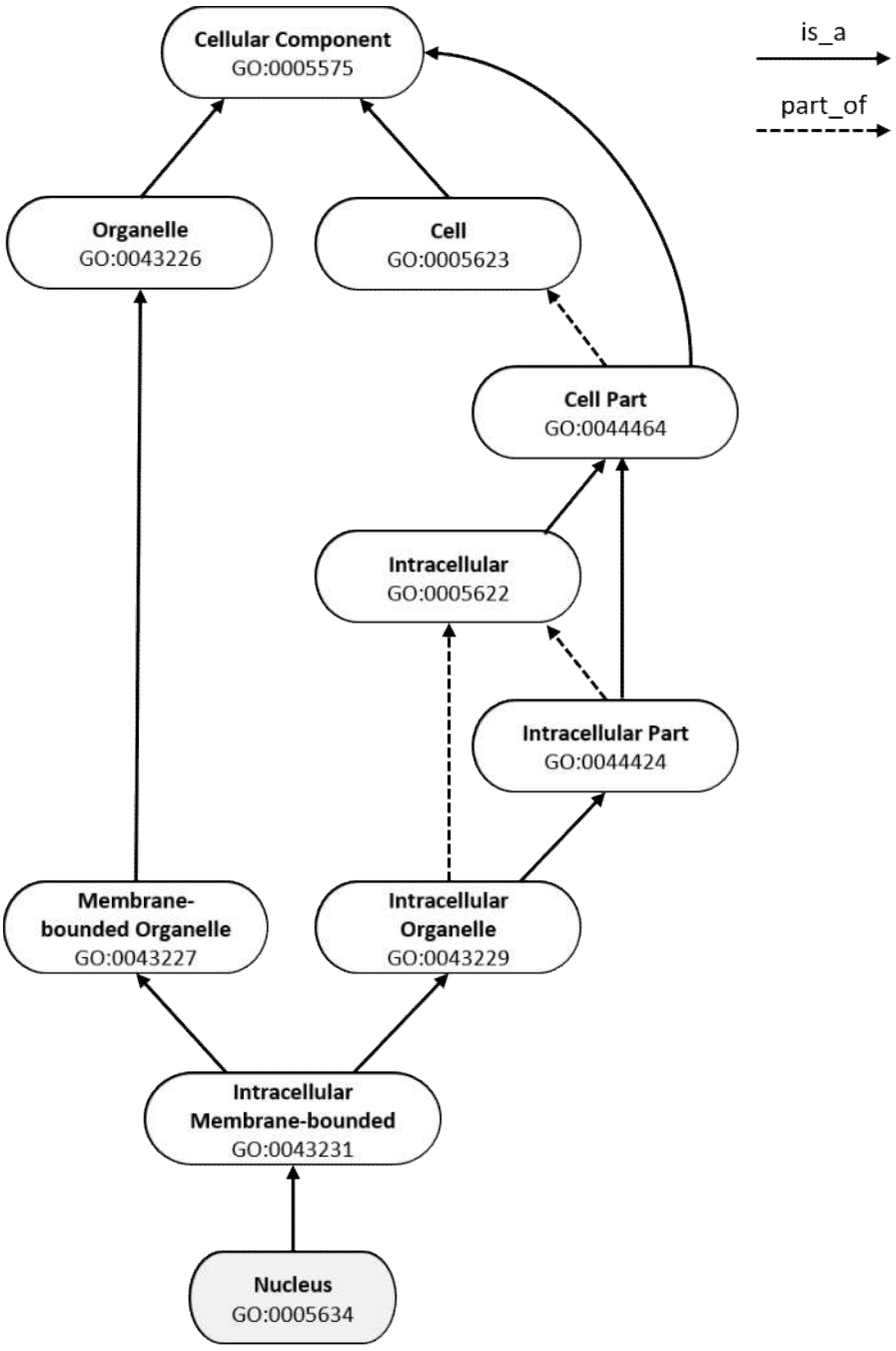

Figure 2.3: Excerpt from the GO graph, referring to the class GO:0005634 (representing the concept "Nucleus"), and including its ancestors [22].

Probably the most important and matured application case of Semantic Web and ontology-based approaches is the area of life sciences, biotechnology, medicine, and pharmaceutical research and development [18]. In the last few decades, novel genomic technologies have greatly expanded the volume and detail of biomedical data, underlining the need to improve its management, integration, and analysis [9].

The development of the Gene Ontology (GO) in 1998 presented the first instance of ontologies being used in this context. Since then, new biomedical ontologies have provided knowledge representation for several other domains, ranging from medications and diseases, to disease processes and genes [23]. The inherent complexity of these subjects also means that biomedical ontologies have certain attributes not usually found in other kinds of ontologies, including [23, 24]:

- **Large size** - Many of the most used biomedical ontologies have tens or even hundreds of thousands of classes, and handling them can be computationally challenging.
- **Lexical richness** - Biomedical ontologies possess a rich and complex vocabulary, where classes are typically described by several annotations, such as labels and synonyms.
- **Particular semantics** - Biomedical ontologies have few properties and relatively simple semantics.
- **Modeling differences of the same domains** - While it is common for distinct ontologies to model the same domains differently, the complexity of the biomedical domain makes these changes more profound.
- **Susceptibility to modeling mistakes** - The large and intricate nature of biomedical ontologies leaves them especially prone to modeling errors. This has also led to difficulties in their development.

The success and continuous use of biomedical ontologies is reflected in their ever increasing size and number. Amid their proliferation, finding a specific ontology within a given domain is a task made easier when these are stored within a single source. This is where ontology repositories come into relevance. The largest of these collections is the Bioportal repository, developed by the National Centre for Biomedical Ontology (NCBO), and currently containing over 700 ontologies available in a variety of formats, such as OWL an OBO [9]

Furthermore, the evolution of biomedical data means that ontology vocabulary is constantly expanding. Consequently, there is an increasing amount of overlapping and conflicting content between biomedical ontologies, which needs to be addressed when considering the interoperability between ontology-based systems. Ontology matching is a fledgling research field concerned with achieving a better integration of ontologies by developing methods that map, or connect, semantically related concepts between them [25].

In healthcare, ontologies play an important role through a variety of applications. They can be used to annotate medical datasets, literature and patient records in order to better access data and extract knowledge. By providing a common terminology for biomedical entities, they can also standardize and integrate of data from different sources, contributing to the interoperability between ontology-dependent medical systems. As a source of computable domain knowledge, they have also proven useful for natural language processing (NLP) and clinical decision support systems. Otherwise, they are critical to hypothesis generation and knowledge discovery in a data driven approach to biomedical research [18, 26].

2.1.2 Semantic Annotation of Data

As seen before, the value of any kind of data is greatly enhanced when it exists in a form allowing integration with other data [16]. One way to achieve this is using ontology-based semantic annotations to associate entities in text with ontology classes (Figure 2.4). By annotating multiple bodies of data with the same ontology vocabulary, these can be standardized under a common format, exchanged, and combined in other processing tasks [26].

Moreover, information retrieval from annotated data boasts higher recall and precision. Because ontologies are organized in a hierarchy, annotation queries can be expanded to the descendants of the original input term, while also being enriched with synonym concepts. For example, Figure 2.4 shows how more detailed symptomatic data can be added to simple trait annotations.

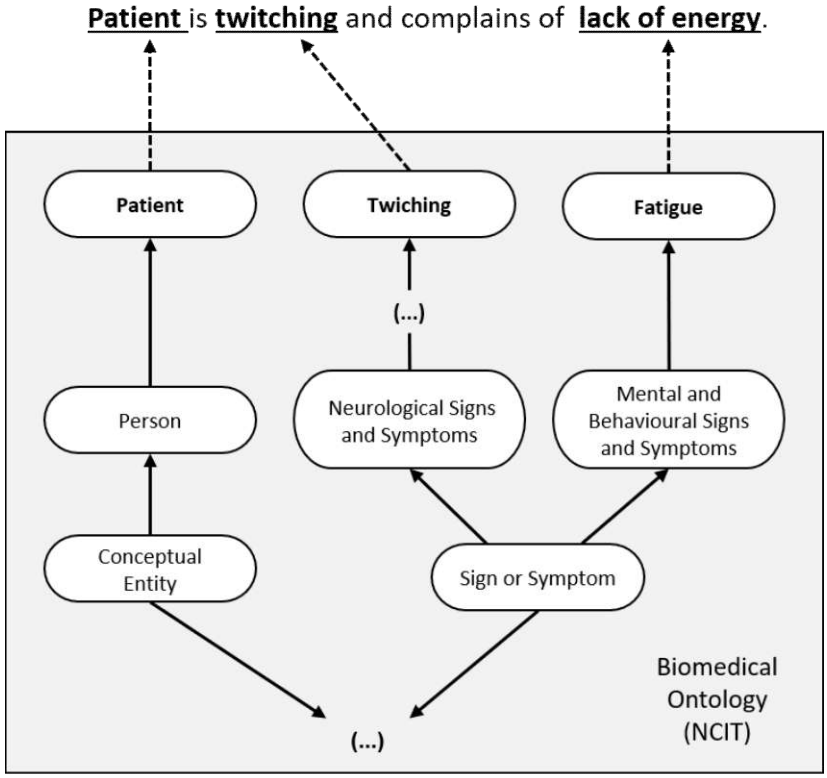

Figure 2.4: Example of a semantic annotation, where concepts from the NCIT ontology are matched to terms in a sentence.

The process of manually annotating each term in a document is laborious, expensive, and altogether too inefficient to be used on large-scale and biomedical data. Semantic annotator software remedies this by automating the extraction of terms from text sources, and disambiguating them against ontology concepts [2, 17]. Many annotator tools have been developed within and out specific terminologies, but have been particularly useful in the biomedical domain. Some of the most noteworthy are described below:

- **MetaMap** - One of the best known biomedical annotators, it maps biomedical entities within an input text with corresponding concepts in the Unified Medical Language System (UMLS) Metathesaurus. MetaMap offers limited configuration options to fine-tune the annotation process, and can return a matching score for every annotation. However, its underperforms when annotating large corpora or when disambiguating ambiguous terms [2, 27].
- **NCBO Annotator** - Another well known annotator that uses biomedical ontologies and thesauri (e.g. Bioportal and UMLS) to annotate biomedical concepts within an input text. Unlike other annotators, it can extend direct annotations with semantically related concepts from other relevant ontologies and/or from ontology mappings. Because it is available as a free Web service, it can be easily integrated in current programs and workflows [28].
- **ConceptMapper** - A general purpose dictionary look-up tool, developed as a component of the open-source UIMA NLP framework. Several mapping options give it a flexibility in concept detection that other annotators do not possess, though finding the ideal configuration can be a challenge. While generic in application, it has outperformed state-of-the-art biomedical annotators [2].
- **BioMedical Concept Annotation System (BeCAS)** - An open-source Web-based tool for the semantic annotation of biomedical texts, primarily biomedical research papers. It adopts a modular system approach, containing separate functionalities for text pre-processing and concept detection. Terms are mapped to concepts on a custom database compiled from various meta-thesauri and ontologies [2].

2.2 Semantic Similarity

The properties of ontologies and ontology-based semantic annotations can be exploited for various data mining and analysis tasks. Specifically, and most relevant to this work, is the use of semantic similarity measures to compute semantic similarity between data items.

Semantic similarity can be understood as the strength of semantic interactions between two elements when considering their taxonomic relationships. These elements can be compared in regard to the constitutive properties they share and those which are specific to them [29], using semantic similarity measures, i.e., functions that, given two ontology terms or two sets of terms annotating two entities, return a numerical value reflecting the closeness in meaning between them [12]. There is a wide array of implementations for similarity measures, many of which are here compiled in a more comprehensive format in Table 2.1, adapted from Guzzi et al. [30].

So far, semantic similarity measures have been used in important machine learning applications, such as computational prediction of protein-protein interactions, disease diagnosis from patient phenotype data, or in classification of chemicals from their structural information [9].

2.2.1 Term Semantic Similarity

Term semantic similarity, or pairwise similarity, is measured between two ontology concepts. Most similarity measures are of this type [29]. For example, Resnik [31] proposed a pairwise, node-based measure comparing two concepts c_1 and c_2 according to the information they share, formally:

$$sim_{Resnik}(c_1, c_2) = IC(MICA(c_1, c_2)) \tag{2.1}$$

The information content (IC) of a concept c is a reference to its specificity, which is regarded as the amount of information the concept conveys [29]. It can be quantified as the the negative the log likelihood $-log(p(c))$, where $p(c)$ is the probability of it occurring in a given corpus. Resnik originally defined $p(c)$ as:

$$p(c) = \frac{\sum_{w \in W(c)} count(w)}{N}, \tag{2.2}$$

where $W(c)$ is the set of words in the corpus subsumed by c, and N refers to the total number of nouns observed (excluding those not subsumed by any class). Other measures expand on this notion. Sanchez et al. [32] gave a more updated definition of IC:

$$p(c) = \frac{\frac{|leaves(c)|}{|subsumers(c)|} + 1}{max\_leaves + 1}, \tag{2.3}$$

where $leaves(c)$ is the set of leaves of c, $subsumers(c)$ is the set of ancestors of c, including itself, and max_leaves is the total number of leaves in the taxonomy. Regardless of the IC measure, pairwise similarity measure in (2.1) is a function of the IC of their common ancestor

which maximizes the IC metric, i.e, the Most Informative Common Ancestor (MICA) [29].

2.2.2 Entity Semantic Similarity

Entity Semantic Similarity, or groupwise similarity, is measured between objects as sets of concepts [29]. The maximum (MAX), average (AVG), and best-match average (BMA) [33] are three notable examples of measures employing aggregation strategies, which summarize pairwise similarity scores between each concept of a different set. From two annotated entities A and B, the MAX approach finds the maximum similarity between each term in A and B:

$$sim_{MAX}(A, B) = MAX_{t_1 \in A, t_2 \in B}(sim(t_1, t_2)) \tag{2.4}$$

The AVG approach is given by the average similarity between each term in A and each term in B:

$$sim_{AVG}(A, B) = AVG_{t_1 \in A, t_2 \in B}(sim(t_1, t_2)) \tag{2.5}$$

The BMA approach is given by the average similarity between each term in A and its most similar term in B, averaged with its reciprocal to obtain a symmetric score [34]:

$$sim_{BMA}(A, B) = \frac{AVG_{t1}(MAX_{t2}sim(t1, t2)) + AVG_{t2}(MAX_{t1}sim(t1, t2))}{2}, t_1 \in A, t_2 \in B \tag{2.6}$$

On the other hand, SimGIC is a graph-based groupwise measure that directly computes the similarity of A and B through the sum of the IC of each term in the intersection of A with B divided by the sum of the IC of each term in their union [34], formally:

$$simGIC(A, B) = \frac{\sum_{t \in \{A \cap B\}} IC(t)}{\sum_{t \in \{A \cup B\}} IC(t)} \tag{2.7}$$

Table 2.1: Summary of Term Semantic Similarity Measures [30].

Type	Name	REF	Term IC	MICA	ACA	Path Length	Term Depth	VSM
Groupwise	Ali and Deane	[35]	No	Yes	No	No	No	No
	Cho	[36]	Yes	Yes	No	No	No	No
	Cosine	[37]	No	No	No	No	No	Yes
	Czekanowski-Dice	[38]	No	No	Yes	No	No	No
	Dice	[37]	No	No	Yes	No	No	No
	FMS	[39]	Yes	No	No	No	No	No
	IntelliGO	[40]	Yes	Yes	No	Yes	Yes	Yes
	Jaccard	[37]	No	No	Yes	No	No	No
	Kappa statistics	[41]	No	No	Yes	No	No	No
	NTO	[42]	No	No	Yes	No	No	No
	PL	[43]	No	No	No	Yes	No	No
	simGIC	[34]	Yes	No	Yes	No	No	No
	simLP	[44]	No	Yes	No	No	Yes	No
	simNLP	[45]	No	Yes	No	No	Yes	No
	simUI	[44]	No	No	Yes	No	No	No
	SSA	[46]	Yes	Yes	Depends on Measure Used			
	TO	[47]	No	No	Yes	No	No	No
	TAS	[48]	No	Yes	No	No	No	No
	Weighted cosine	[49]	Yes	No	No	No	No	Yes
	WJ	[37]	Yes	No	Yes	No	No	No
Pairwise	Annotation cosine	[50]	No	No	No	No	No	Yes
	G-SESAME	[51]	No	No	Yes	Yes	No	No
	GraSM	[52]	Yes	Yes	No	No	No	No
	Jiang and Conrath	[53]	Yes	Yes	No	No	No	No
	Lin	[54]	Yes	Yes	No	No	No	No
	Othman	[55]	Yes	Yes	No	Yes	Yes	No
	PS or PK-TS	[56]	No	Yes	No	Yes	Yes	No
	Resnik	[31]	Yes	Yes	No	No	No	No
	RSS	[57]	No	Yes	No	Yes	Yes	No
	SB-TS	[58]	No	No	No	No	Yes	No
	simIC	[59]	Yes	Yes	No	No	No	No
	simRel	[33]	Yes	Yes	No	No	No	No
	SSM	[60]	Yes	Yes	No	Yes	Yes	No
	TCSS	[61]	Yes	Yes	No	No	No	No
	Wu	[62]	No	No	Yes	No	No	No
	Wu-Palmer	[63]	No	Yes	No	Yes	Yes	No
	XOA	[64]	Depends on Measure Used					Yes

Columns Term IC, Some common ancestors (MICA), All common ancestors (ACA), Path length, Term depth and VSM refer to the features of the measures described in the text. NTO, normalized term overlap; PL, path length; PS or PK-TS, pekar-staab term similarity; SSA, semantic similarity of annotations; TO, term overlap; TAS, total ancestry similarity; WJ, weighted Jaccard; XOA, cross ontological analysis

2.2.3 Semantic Similarity with Multiple Ontologies

The increased availability of research data has made it possible for several biomedical domains to be converted into ontologies, and then used as background knowledge to annotate and compute semantic similarity on data of proportionally diverse kind [32]. But despite their abundance, each individual ontology is still by definition constricted to a specific scientific field.

Consequently, when the analyzed documents integrate feature data from several and often unrelated subjects, using a single ontology for semantic similarity tasks means that only a fraction of the data can be taken into account, leaving out the terms which are not matched by any of the ontology's concepts [32]. To solve this coverage problem, recent approaches have opted to use multiple ontologies as background for semantic similarity, so that:

- The content of one ontology can complement what another one lacks, i.e., both can cover more concepts and supply additional information.
- When the content of two ontologies does overlap, mistakes can be corrected, and misconceptions of one and the same concept can be filtered and minimized.

This enables the comparison of elements that are not formally defined in the same ontology, while also refining the whole process by incorporating a larger amount of information [29]. Multiple ontology similarity already shows promising prospects for health applications, including in novel methods for patient classification and stratification, and the analysis and mining of large-scale patient data [9].

Unfortunately, the integration of multiple ontologies into a single semantic similarity measure is a less practical problem, since joining ontologies into a single structure must necessarily deal with inconsistencies and mismatching concepts between them [32, 65]. Current approaches have dealt with this by either designing new similarity measures that can handle multiple ontologies, or by lifting existing single-ontology measures into multi-modal measures [13].

2.2.4 Semantic Similarity Measure Software

With semantic similarity measures being able to objectively quantify similarity, a new focus has also shifted towards developing software capable of adapting, or extending measures into a framework that can automatically compute semantic similarity between ontology concepts or annotations. Several new tools have been implemented for this purpose. Some are dedicated to a specific ontology and terminology, while others opt for a generic approach, supporting many ontology formats and domains. Below are just a few examples with a brief description.

- **SimPack -** Java library adapting similarity measures for pairwise comparison of ontology concepts. SimPack is a generic application, and may be used with ontologies of either OWL or RDF formats [66].
- **Similarity Library -** Java library originally developed for accessing the WordNet lexical database, now extended to support MeSH and the Gene Ontology. It implements semantic measures to find both pairwise and group-wise similarity of concepts [67].
- **DOSim -** R package dedicated to the Disease Ontology, and used to compute the similarity between diseases and to measure the similarity between human genes in terms of diseases. It also incorporates a DO-based enrichment analysis function that can be used to explore the disease feature of an independent gene set [68].
- **Semantic Measures Library (SML) -** A generic open source java library and command line tool, which compiles a large collection of knowledge-based similarity measures for both group-wise and pairwise comparison of concepts/annotations. These measures are not specific to any ontology language or domain, enabling SML to deal in various knowledge representation formats (e.g. OBO, OWL, and RDF). SML also enables large scale computation and analysis of semantic measures, supporting multi-thread processes for fast parallel computation [29]

2.3 Clustering

Clustering, or cluster analysis, is a statistical technique that helps reveal hidden structures in data by grouping entities or objects with similar characteristics into homogeneous groups, while maximizing heterogeneity across different groups [69]. It is used in data mining for making predictions based on groups, hypothesis generation, data exploration, and data reduction, i.e., grouping similar entities into homogeneous classes to organize large quantities of information. In biomedicine, it has been applied to characterize psychiatric patients on the basis of clusters of symptoms, to find groups of genes with similar biological functions, or identify medical patient groups most in need of targeted interventions [70].

There are many kinds of clustering methods, but the most commonly discussed distinction among them is whether the set of clusters is nested or unnested, or in more traditional terminology, hierarchical or partitional [69]. In detail:

- **Partition methods** work by first partitioning a set of n objects into k clusters, with $k \leq n$, and then iteratively moving objects from one group to another until an optimum is reached for an objective criterion, such as a dissimilarity function based on distance, so that the objects within a cluster are "similar" to one another and "dissimilar" to objects in other clusters in terms of the dataset attributes.
- **Hierarchical methods** create a hierarchical decomposition of a given set of data objects. This can be done taking a agglomerative approach, which starts with each object forming a separate group. It then successively merges objects or groups close to one another, until all the groups are merged into one (the topmost level of the hierarchy), or a termination condition holds. Alternatively, a divisive approach starts with all the objects in the same cluster. In each successive iteration, a cluster is split into smaller clusters, until eventually each object is in one cluster, or a termination condition holds [11].

Clustering was used in this work to join patients into groups using biomedical text and ontological data. While hierarchical clustering is known to produce results more expediently, it is outperformed by partition clustering for larger datasets [70], which, as discussed before, are more typically found in biomedical fields. Hence, this work proceeded using the latter method. More specifically, it used the popular K-means algorithm, a fundamental partition method used for small to medium-sized data. However, when data becomes larger, or when clusters take a more complex shape, K-means can be extended into other techniques, like spectral clustering [11].

2.3.1 K-Means

K-means is a very simple and appealing partition method that represents a cluster as its centroid, i.e. its centre point. Given a dataset C, it first clusters data objects into k clusters, by arbitrarily selecting k objects as centers, and assigning the remaining objects to their nearest centre (see Figure 2.5-a). These two steps are then repeated, updating the cluster centres and redistributing objects until there is no difference in object assignment.

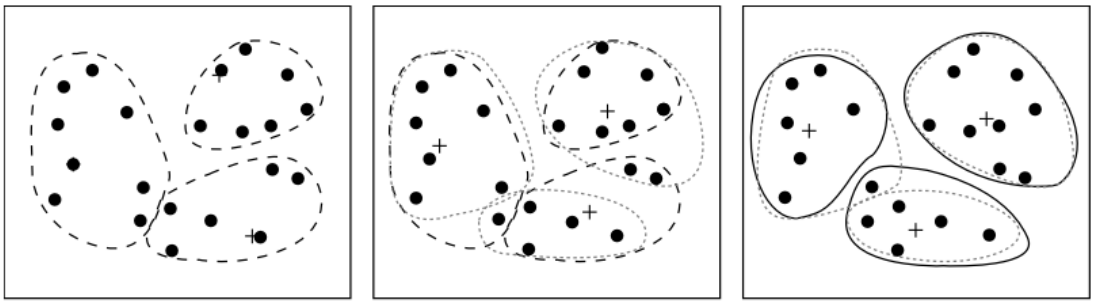

Figure 2.5: Clustering of a set of objects using the k-means method; for (b) update cluster centers and reassign objects accordingly (the mean of each cluster is marked by a +) [11].

It is therefore the goal of K-means to minimize the distance between any object x_i and a cluster centre c_i, keeping clusters as compact as possible. This distance, written as $dist(x_i, c_i)$, usually refers to the euclidean distance, and can be used to quantify the quality of the cluster through the within-cluster variation, defined as:

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} dist(x_i, c_i)^2 \tag{2.8}$$

Unfortunately, while K-means is a fast and practical approach to clustering, it is not the most accurate, nor is it guaranteed to find a global optimum when clustering C [11]. Because results can depend on the initial cluster partition, it is common to run K-means multiple times with different initial cluster centres to improve the outcome. Alternatively, there are also variants of K-means that change how this step is processed [11]. For example, the Python Scikit Kmeans algorithm used in this work implements K-means++ as a default initialization method. It works by choosing the first center uniformly at random, and then selecting a new centre from an object $x_i \in C$ until reaching k centres, each with the probability:

$$\frac{D(x)^2}{\sum_{x \in C} D(x)^2} \tag{2.9}$$

Where $D(x)$ is the shortest distance from a data point to the closest center which was previously chosen [71]. The rest of the algorithm proceeds as k-means, but this step increases the quality of the final clustering, while also speeding up its computation [11].

2.3.2 Spectral Clustering

Sometimes, data objects are described by hundreds of features, or dimensions, and the sheer size of this high-dimensionality data makes it harder for standard clustering methods to mine underlying patterns. Dimension reduction techniques solve this by removing or extracting features in order to obtain a reduced representation of the dataset that is smaller in volume, but maintains the integrity of the original data. However this comes at the cost of loosing data, which can reduce the chances of detecting more subtle clusters [11].

On the other hand, there are other clustering algorithms more suited for this kind of data. Spectral clustering (Figure 2.6) attempts to find clusters in subspaces (i.e., a derived space) of the original data, without ever modifying it. Consider a set of data objects $x_1, ..., x_n$ with a similarity $s_{ij} \geq 0$ computed between every distinct pair. Objects can instead be represented on a graph $G = (V, E)$, where $V = \{v_i, ..., v_n\}$, and where each vertex vi corresponds to the object xi. G is weighted, so that the edge between two vertices v_i and v_j has a weigh $w_{ij} \geq 0$. If $w_{ij} = 0$, then the two points are not connected. Using these weights, an adjacency matrix (or affinity matrix) A, can be derived from G. Formally:

$$W = (w_{ij})_{i,j=1,...n} \tag{2.10}$$

Furthermore, a degree matrix D, a diagonal matrix based on degrees of vertices in G, is built. The degree of a vertex $v_i \in V$ is the number of the edge endpoints to v_i, defined as:

$$d_i = \sum_{j=1}^{n} w_{ij} \tag{2.11}$$

From W and D, a third matrix, a Laplacian L, can be calculated. Spectral Clustering relies on the properties of Laplacians to find eigenvectors and their respective eigenvalues. The eigenvectors of a square matrix are the nonzero vectors that remain proportional to the original vector after being multiplied by the matrix. Mathematically, a vector v is an eigenvector of L, if $Lv = \lambda v$, where λ is the eigenvalue [11]. There is more than one way of finding L. Bellow, (2.12) and (2.13) are, respectively, an unormalized and normalized Laplacian:

$$L = D - W \tag{2.12}$$

$$L_{syn} = D^{-\frac{1}{2}} L D^{\frac{1}{2}} = 1 - D^{\frac{1}{2}} W D^{\frac{1}{2}} \tag{2.13}$$

The first k eigenvectors correspond to the k smallest eigenvalues. A k-means clustering algorithm can be applied to a matrix U containing the vectors $v_1, ..., v_k$ as columns, where each clustered point corresponds to a row of U. Clusters can then be projected back to the original data [72, 73]. The dimensionality of the new space is set to the desired number of clusters. This setting expects that each new dimension should be able to manifest a cluster [11].

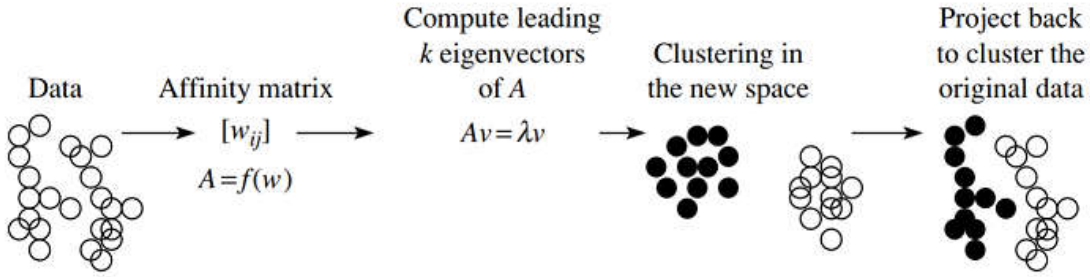

Figure 2.6: The framework of spectral clustering approaches [11].

Spectral clustering is a simple and efficient technique, with better performance than other more traditional approaches. Unlike K-means, it does not assume that the clusters will take a convex shape, as is able to solve very general problems like intertwined spirals. However, its worth noting that there are scalability issues, since computing eigenvalues over a large matrix is computationally expensive [11].

2.3.3 Data Mining Tools

- **Python Modules**

  Python has a large and active ecosystem of third-party packages that provide implementations of several data mining functionalities [74]. Some of these, all open-source, are:

  (a) **Pandas** - Is built on - and improves upon - the NumPy library , which provides Python with efficient storage and manipulation of dense typed arrays. Panda's objects, in particular the DataFrame, can be though of an enhanced version of numPy's own structured arrays, enabling convenient storage interface for labeled data, as well as implementing a number of powerful data operations [74].

  (b) **Scikit-Learn** - Provides state-of-the-art implementations of a range of common machine-learning algorithms, and is characterized by a clean, uniform, and streamlined API, as well as by very useful and complete online documentation. The package relies on the NumPy library for data handling, and SciPy for efficient algorithms. It is mostly written in Python, but it also incorporates the LibSVM and LibLinear C++ libraries, and while being a collaborative effort, it boasts of consistent and reliable implementations over feature quantity [74, 75].

  (c) **Mlxtend** - Another open-source library containing a variety of core algorithms and utilities for machine learning and data mining, focusing on user-friendly and intuitive APIs and compatibility to existing machine learning libraries, such as Scikit-learn [76].

  (d) **Matplotlib** - A multi-platform data visualization library built on NumPy arrays, and designed to work with the broader SciPy stack. One of Matplotlib's most important features is its ability to play well with many operating systems and graphics backends [74].

- **R and RStudio**

  R is a free tool designed for statistical analysis, visualization and reporting. Its source code is written in R itself, but also receives a heavy contribution from C and FORTRAN. R has many properties that most other languages do not have, such as allowing the state of objects and results to be seen one command at a time without previous compiling, and having powerful data structures, namely the data.frame, which can handle mixed data types in a spreadsheet-like format. Similarly to python, the biggest reason for R's popularity is its collection of user contributed packages, offering many options for data mining algorithms and libraries for further data manipulation [77].

  RStudio is an open-source IDE for R that facilitates its use through useful features, including code completion, execute from source, and searchable history [77, 78].

2.4 Clustering with Semantic Similarity

Distance based clustering algorithms find associations between objects by comparing the data that describes them following a distance metric, such as Euclidean distance for numerical attributes, or Levenshtein distance for categorical ones [11]. However, typical metrics fail to recognize semantic relations in unstructured/textual data. Take for instance the two strings:

a) "Large intestine adenocarcinoma, metastatic to liver".

b) "Adenocarcinoma of colon, with hepatic metastasis".

Both sentences are semantically identical since they refer to the same conceptual medical idea, but because they do not share complete commonality from a lexical point of view, they are placed at a distance by categorical measures. This limitation has been addressed by integrating ontologies and their formal semantic definitions of concepts, which are made available to data mining techniques through the semantic annotation of textual content.

Semantic similarity can thus look to ontologies to provide a more realistic measure of object closeness, and although its use in clustering is still relatively recent, it has enjoyed significant success in various fields of scientific research. Table 2.2 presents a summary of the most relevant applications.

Table 2.2: Some applications of semantic similarity for clustering.

Ref	Domain	Pairwise Measures	Groupwise Measures	Clustering Algorithm
[79]	GO	Lin, JC, Resnik, simRel	MAX, AVG, simGIC, Cosine Czekanowski-Dice, Kappa	Aglomerattive Hierarchical Clustering
[80]	GO	GCSM	IntelliGO	Hierarchical and Fuzzy Clustering
[81]	GO	Lin	TO, BMA, simGIC	K Nearest
[82]	GO	*	-	K-Medoids
[83]	WordNet	-	Cosine	Bisecting K-means
[84]	WordNet	-	Cosine	K-means
[85]	WordNet	JC	Average	Spectral Clustering
[86]	HPO	Resnik	BMA	PAM

*unspecified; BMA, Best Match Average; GCSM, Generalized Cosine-Similarity Measure; GO, Gene Ontology; HPO, Human Phenotype Ontology; JC, Jiang and Conrath; Kappa, Kappa Statistics; PAM, Partitioning Around Medoids; TO, Term Overlap;

Semantic similarity has been instrumental to document clustering in tackling problems of synonymy, ambiguity and lack of a descriptive content. Bouras et al. [84] extracted words from each of several news articles and enriched its most frequent terms with WordNet hypernym subgraphs. K-means clustering of articles used the cosine similarity measure and revealed an increase in cluster quality using this approach. Desai et al. [83] used WordNet for word sense disambiguation (WSD), but selected from multiple possible word senses whichever maximized pairwise similarity with it, replacing words with synset IDs and calculating cosine similarity between documents for Bisecting K-means clustering. Blokh et al. [85] clustered social media news messages by theme, calculating sentence closeness as a function of pairwise similarity between their words as WordNet synsets, and using Spectral Clustering with a pre-computed sentence similarity matrix.

The growing size of gene data has also made it a target of new similarity assisted clustering applications. Ovaska et al. [79] developed a methodology and software that performed hierarchical clustering on differently expressed genes (DEGs) from microarray data, based on functional GO annotation and calculated similarity scores between genes. The authors benchmarked a variety of existing similarity measures to test speed gain, and demonstrated its real-world application. Benabderrahmane et al. [80] used a likewise method with their own IntelliGO similarity measure on groups of DEGs with similar expression levels, to identify subsets of genes displaying consistent expression and functional profiles. Yu et al. [81] calculated similarity between GO-annotated genes of two species with homology, and determined their k nearest neighbors to replenish annotations for incompletely annotated genes.

Also, Westbury et al. [86] annotated a large collection of BPD patients with the Human Phenotype Ontology (HPO), measured their semantic similarity, and clustered patients with Partitioning Around Medoids (PAM) algorithm. This approach successfully found individuals with phenotypic similarities, who are more likely to have causal genetic variants in the same or related genes.

More recently, Ostaszewski et al. [82] clustered Parkinson's and Alzheimer's disease maps (i.e., diagrams depicting pathway interactions among disease-related bioentities) as graphs, using the GOSemSim R package to calculate pairwise similarity between term elements of the maps within GO, and a K-medoid based clustering algorithm.

While these works are still limited to very few ontology domains, all underlined the benefits gained from the inclusion of an external knowledge source. In spite of this, and as far as this work is concerned, there have been no application instances using more than a single ontology for clustering, which would have given the prospects for a more encompassing and accurate mining.

3 Methodology

This work aimed to develop a novel methodology to analyze and cluster patient data possessing of a strong textual/nominal component, by relying on semantic similarity scores as a metric for distance-based clustering approaches.

This chapter is organized as follows. Section 3.1 gives a general overview of Semantic Similarity Clustering. Sections 3.2 to 3.5 offer a closer look into the methodology, covering each step of in greater detail. Section 3.6 explains how semantic similarity can be used to interpret and semantically describe clusters.

3.1 General Architecture

The methodology proposed in this work can be generalized to any biomedical entities which can be semantically annotated with existing ontologies. Its main architecture is represented in Figure 3.1. Fundamentally, it uses single or multiple ontologies as a source of semantic knowledge to annotate terms in textual patient data. Then, understanding each patient as an annotated object, it computes their semantic similarity scores. These scores are arrayed into a matrix, and inputted as a kernel for different clustering strategies. The resulting clusters are studied through a semantic description of their content, supporting a more accurate and informed extraction of knowledge. This process is carried out in a series of steps:

1. **Ontology Selection** - Finding an ontology, or combination of ontologies, that is best suited to recognize the specific vocabulary found in patient data, by means of measuring how many and accurate its annotations can be, and ranking its performance against other options;
2. **Semantic Annotation** - A controlled annotation of patient data with the previous ontology selection, able to filter unwanted annotations or kinds of annotations, while also optimizing the process for larger-scale application;
3. **Semantic Similarity** - Computing semantic similarity between each unique combination of patients as sets of annotations, using semantic similarity measures;
4. **Clustering** - Clustering patients using their semantic similarity scores as a distance metric for clustering algorithms;
5. **Semantic Description** - Summarizing patient clusters by looking into their collective annotation content, and pinpointing the most prevalent/meaningful concepts. This includes a visualization method to facilitate cluster interpretation.

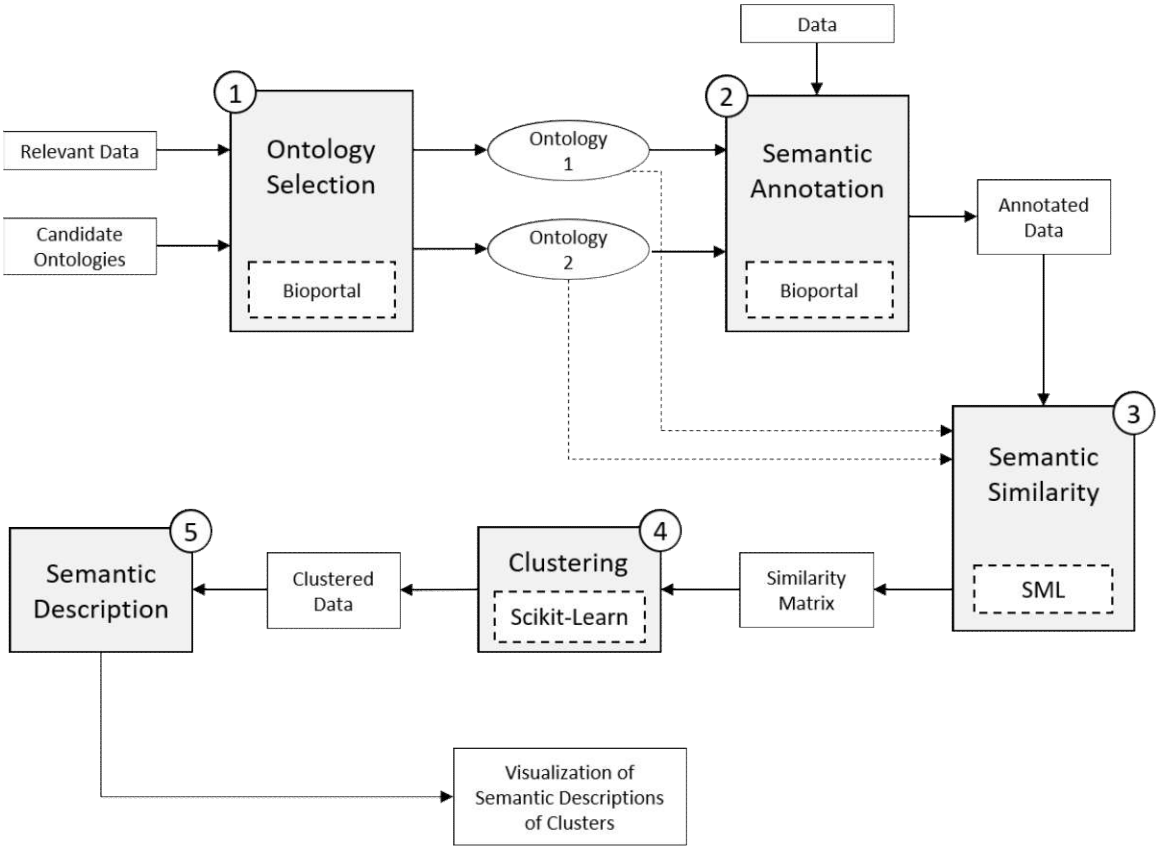

Figure 3.1: Overview diagram of the proposed methodology.

This methodology was here implemented using mostly R scripts in RStudio, which were complemented with applications of Python and Java libraries for more specific tasks, in particular:

- **SML** was used to compute patient similarity scores, as it is extensively documented and tested. It also allows for multiple ontology comparison and has a broader choice of similarity measures, which is all the more important when testing different settings to optimize patient scores and clusters;
- **Scikit-Learn** algorithms were used in clustering steps. Not only it presents extensive documentation and narrative examples, but also provides unique and curated algorithm applications that cover most machine learning tasks.

3.2 Ontology Selection

The process of selecting ontologies prioritizes their capacity to supply formal knowledge for patient data. A possible criteria for this can simply observe the number of terms (i.e., a word or phrase used to describe a thing or to express a concept) it can annotate. However, its also important to consider the accuracy of annotations. For example, if the term "upper motor neuron" is matched with the three concepts "neuron", "motor", and "motor neuron", it cannot be said to have been as accurately annotated as it would have been with the sole concept "upper motor neuron". After all, the idea of a thing is best described by the idea of the class to which it belongs to. This work determined annotation accuracy by how close its class label resembles the matching term or, in other words, its completeness. Three types of annotations are defined, with practical examples of each below in Table 3.1:

- **Complete** - Accounting for the whole term;
- **Multiword** - Accounting for a portion of a multiple-word term;
- **Singleword** - Accounting for a single word of a multiple-word term.

Table 3.1: Examples of annotation completeness

Term	Complete Annotations	Multiword Annotations	Singleword Annotation
"ALS"	Amyotrophic Lateral Sclerosis	-	-
"Upper Motor Neuron"	Upper Motor Neuron	Motor Neuron	Motor;Neuron;Upper
"Tongue Spasticity"	Tongue Spasticity	-	Tongue;Spasticity
"Upper Limbs"	Upper Limbs	-	Upper;Limbs
"High Protein"	-	-	Protein
"Mild Physical Exercise"	-	Physical Exercise	Mild;Physical;Exercise

Hence, the number of terms that can be supplied with complete annotations is taken here as a direct account of the background knowledge potential any given ontology has for patient data. Selecting ontologies can then proceed through the following steps:

1. Perform a manual selection of domain-specific ontologies, i.e., candidate ontologies. This manual selection should take into consideration the domain covered by the data in question;
2. Annotate a representative set of terms extracted from the data with the candidate ontologies;
3. Rank ontologies according to the number of complete annotations they can provide and filter out those unable to provide any complete annotations;
4. Select pairs of ontologies from the ranked list: for each term without a complete annotation in the first ontology, an annotation with the second is attempted.

As discussed before in Section 2.2.3, it is possible to use more than one ontology as a single source of semantic knowledge. Selecting multiple ontologies for annotation should also attempt to maximize the number of completely annotated terms, ensuring a more expansive and realistic semantic representation of data.

3.3 Semantic Annotation Method

Each patient can be modelled as a list (or vector) of terms describing it in the dataset. Since any given term is rarely exclusive to a single patient, it would be redundant, as well as time-consuming, to iteratively annotate over each and every patient list to find the same semantic information. Instead, terms from the patient dataset are uniquely extracted, and annotated once with the selected ontologies, while polling the resulting annotations into a likewise unique set. A matching algorithm can then redistribute annotations from the set over all patient term lists.

The matching process is also subject to criteria, which is based on the overlap between the annotation label and the annotated term. Three matching types are categorized:

- Full Match (FM) - label and term are equivalent (e.g. term “limb” to label “limb”).
- Component Match (CM) - a label is a component of the term (e.g. term “upper limb” to label “limb”).
- Partial Match (PM) - a term is a component of the label (e.g. term “limb” to label “upper limb”).

Match types reflect the complete/multiword/singleword paradigm, as FM emulates complete annotations, and CM connects terms with single and multiword annotations. PM matching acts as a supplementary type, connecting smaller terms with larger annotation labels. Selecting a matching type combination should aim for best possible semantic representation of patients (i.e. FM), but if there are too few annotations available to effectively describe patients, the Matcher’s CM and PM can increase recall at the cost of precision.

3.4 Semantic Similarity

Semantic similarity is computed between patients as groups of annotations, using SML as a source for entity similarity measures, either as standalone functions or through aggregation of term similarity scores. If annotations belong to more than one ontology, SML is also able to extend similarity measures to include them.

3.4.1 Implementation in SML

Access to SML is made available through a command-line tool, compiling various functionalities for semantic analysis, both with domain-specific and generic configurations available, and accepting custom configurations either as individual command parameters, or by compiling these in configuration file with XML format. Alternatively, as a java library, it allows embedding of the source code with the user's own algorithm while enabling other secondary features. Running SML requires certain conditions to be met, namely ensuring its access to:

1. The ontology used to annotate patients, which must be a rooted directed acyclic graph (rDAG), where concepts (or nodes) are linked to other concepts, eventually converging, without self-references, on a single "highest node" (i.e. Root). Implicitly, this means that patients annotated by a certain concept, are also annotated by any other concepts that generalize it.

2. A valid semantic measure. Currently, SML allows for graph-based measures when using ontologies as an external knowledge sources.

3. The actual patient annotations, which must be identified by a valid absolute URI, a unique reference to a concept within the ontology, and possessing correct syntax (e.g. starting with "http://"). Annotations must also be compiled in a specific format (GAF or TSV).

After the patient term lists were converted into annotations lists, these are rearranged into a TSV format (example in Figure 3.2). The path to the ontology files, and the selected semantic similarity measure are all specified within a custom configuration file, ready to be used by SML.

```
PT:Hann081  NCIT:C43248;NCIT:C866;IOBC:20090605632594586O;NCIT:C71386;NCIT:C38311;IOBC:200906082616628286;NCIT:C158551
PT:Hann082  NCIT:C25229;NCIT:C43248;IOBC:200906087944689715;IOBC:20090605632594586O;NCIT:C25189;NCIT:C71404;NCIT:C71405
PT:Hann083  NCIT:C12742;NCIT:C43248;IOBC:200906050983093675;IOBC:200906087944689715;IOBC:201006013384468995;NCIT:C25189
PT:Hann084  NCIT:C25236;IOBC:200906023687363783;IOBC:200906038154517956;NCIT:C43248;NCIT:C54154;NCIT:C13360;NCIT:C12419
PT:Hann085  NCIT:C866;NCIT:C71405;NCIT:C38311;NCIT:C61486;IOBC:200906082616628286;NCIT:C158551;NCIT:C12336;NCIT:C13063
PT:Hann086  NCIT:C25229;IOBC:200906038154517956;NCIT:C53258;NCIT:C38311;IOBC:20090605632594586O;NCIT:C257;NCIT:C13360
PT:Hann087  NCIT:C25229;NCIT:C43248;IOBC:201006013384468995;IOBC:200906085842861324;IOBC:20090605632594586O;NCIT:C25680
PT:Hann088  IOBC:200906038154517956;NCIT:C53258;NCIT:C25189;IOBC:200906082616628286;NCIT:C158551;NCIT:C12336;NCIT:C13063
PT:Hann089  NCIT:C75919;NCIT:C43248;NCIT:C54154;IOBC:20090605632594586O;IOBC:200906022223190030;NCIT:C69165;NCIT:C71404
PT:Hann090  NCIT:C12742;NCIT:C25229;NCIT:C43248;NCIT:C53258;IOBC:20090605632594586O;NCIT:C69165;NCIT:C71386;NCIT:C12419
PT:Hann091  NCIT:C43248;NCIT:C54154;IOBC:20090605632594586O;NCIT:C71386;NCIT:C13360;IOBC:200906087944689715;NCIT:C117252
PT:Hann092  NCIT:C25229;IOBC:200906023687363783;NCIT:C43248;NCIT:C1505;IOBC:20090605632594586O;NCIT:C38311;NCIT:C13360
PT:Hann093  NCIT:C75919;NCIT:C43248;IOBC:200906038154517956;IOBC:200906050983093675;IOBC:20090605632594586O;NCIT:C38311
PT:Hann094  NCIT:C25229;NCIT:C43248;IOBC:200906050983093675;IOBC:201006013384468995;IOBC:200906081125829640;NCIT:C1505
PT:Hann095  NCIT:C12742;NCIT:C25229;NCIT:C43248;IOBC:200906038154517956;NCIT:C54154;IOBC:200906007714277783;NCIT:C12470
```

Figure 3.2: Annotations compiled in the TSV format required by SML. The first column shows patient ID, and the next lists annotations separated by a semi-colon.

Multiple ontology support in SML takes a simple approach, via a re-rooting process that connects the ontologies to a virtual root (Figure 3.3), and integrates them into a single graph. Once this has taken place, similarity can be computed as it would have in single ontology evaluation.

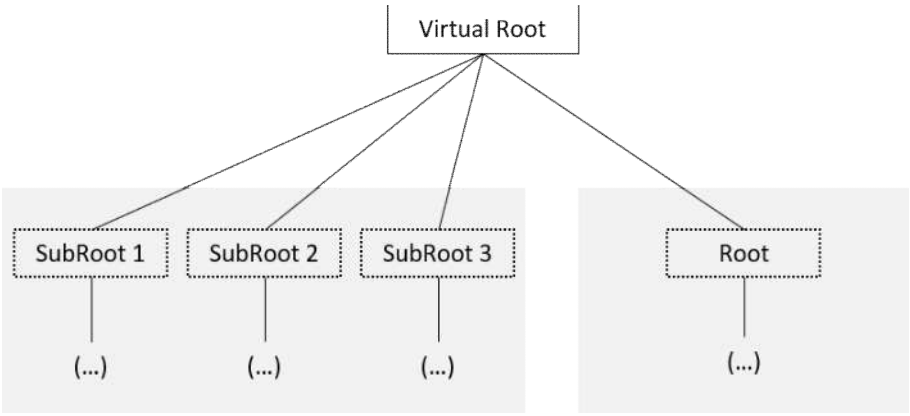

Figure 3.3: Creating a Virtual Root in SML, which connects to the first layer of two different ontologies, whether it is made of a single root or many subroots.

3.5 Semantic Clustering

As seen before on Chapter 2 (Section 2.4), semantic similarity is a valid reflection of object closeness, and has been recently used as a replacement for the distance-based metrics usually applied by clustering approaches. This work adapted the patient similarity scores returned by SML to cluster the patients themselves.

Clustering was done using K-means or Spectral Clustering algorithms from Python's Scikit-Learn module. Both are popular methods with solid implementations, but relying on different methods to find either compact or convex structures hidden in data, which is useful when not having previous knowledge on cluster shape. Patient similarity Scores are accepted in the form of a similarity matrix (example in Figure 3.4), i.e., a precomputed affinity matrix, and inputted as a kernel for clustering.

	ANTA-002	ANTA-003	ANTA-004	ANTA-005	ANTA-006	ANTA-007	ANTA-008	ANTA-009
ANTA-002	1.0	0.379	0.462	0.122	0.357	0.253	0.201	0.201
ANTA-003	0.379	1.0	0.513	0.207	0.385	0.371	0.385	0.300
ANTA-004	0.462	0.513	1.0	0.186	0.660	0.537	0.350	0.273
ANTA-005	0.122	0.207	0.186	1.0	0.314	0.281	0.228	0.228
ANTA-006	0.357	0.385	0.660	0.314	1.0	0.623	0.420	0.328
ANTA-007	0.253	0.371	0.537	0.281	0.623	1.0	0.503	0.503
ANTA-008	0.201	0.385	0.350	0.228	0.420	0.503	1.0	0.414
ANTA-009	0.201	0.300	0.273	0.228	0.328	0.503	0.414	1.0

Figure 3.4: Example of a similarity matrix, compiled from a list of semantic similarity scores.

3.6 Semantic Description of Clusters

After clustering, the next step in knowledge extraction is understanding the results. Clusters can be interpreted in regards to the objects populating them, by finding and isolating data that is more specific to one subset than to all others, i.e., data with the highest variance. Some feature selection techniques already employ this notion [11], but since patient clusters are based on annotations rather than direct feature data, they cannot be used.

Instead, this work proposes an novel approach to describe patient clusters in deference to their most relevant annotations. A semantic description of clusters will aim to further integrate external ontology data into cluster analysis, by linking patient annotations to their concepts in an ontology, and extending them with their ancestors. For example, consider a cluster populated with n patients. Let S be a set of concepts annotating a single patient inside the cluster, so that S is a semantic abstraction of the patient. Using the OWL API java library, a list of ancestors for every concept in S can be retrieved from the ontology, and uniquely added to S, effectively modelling the patient as a sub-graph within the ontology. This process is repeated through every patient in the cluster, generating a list $L = \{S_1, \ldots, S_n\}$ of as many sets as there are patients.

Hence, every cluster can be interpreted, through its own L list, as a series of overlapping subgraphs, and a scoring function can then determine which concepts are the most significant. Two cluster description measures were defined in this work:

1 Representativity Score

For every distinct concept $c \in L$, a representativity score R can be calculated as a compromise between concept presence in cluster, and overall semantic relevance. Formally:

$$R = F \times IC \tag{3.1}$$

Where F is the frequency of c in the cluster, i.e., the number of times it occurs in L. The IC refers to the Information Content of c in its ontology, which is found once again by using OWL API with a valid IC measure. Calculated R scores are normalized between $[0, 1]$.

2 Annotation P-values

This measure was based on gene set enrichment analysis (GSEA), a statistical test used in gene-expression analysis. GSEA finds an enrichment score for a set of genes by ranking it against a larger list of genes, and measure over-representativity [87]. Using this approach, concepts specific to a given cluster, and therefore inside L, are compared against a larger list A of the concepts found in every cluster, by walking through A and increasing a running-sum statistic if the concept is encountered in L, and decreasing it otherwise. A final P-value for statistical significance is then obtained through a permutation test.

The R package ClusterProfiler was used to compute a P-value for very distinct $c \in L$. It is a Bioconductor tool built for the statistical analysis of functional profiles for gene clusters, but with a generic application accepting custom annotations [88].

This approach, while easier to implement, and less time-consuming to process than if using R-scores, does not take directly into account the IC of the term, disregarding semantic context.

After obtaining an initial cluster description, a refinement process can manually filter out specific concepts in L before calculating scores. However, some automatic filters were implemented to remove redundant concepts and decrease the computational load, targeting concepts:

- Below an IC threshold, avoiding concepts which are too vague to transmit actual meaning (e.g. the root).
- Present in all clusters, with presence above 50%, where presence is the ratio between the concept's F in a cluster and n.
- Which are direct ancestors of other concepts in the same S, both having the same F value.
- Having the same label as another concept in S from a different ontology used to annotate the patients (if using multiple ontologies). A list of concepts in common between ontologies is obtained using the AgreementMakerLight (AML)[89] ontology matching system . The concept with the lowest score is eliminated.

3.6.1 Output

The output of the description consists of a ranked list of annotations for every cluster of patients. Each annotation is displayed with information concerning its corresponding class identifier and label, its frequency on the clusters's patients, its IC, the calculated R-Score and P-value, and the label of the cluster it describes. Table 3.2 shows an example of a semantic description over a set of patient clusters. Analysis can be simplified by shortening the list to the mot concepts with the highest R-Score, or lowest P-Value.

Table 3.2: Example of a semantic description of three clusters.

Class ID	Label	Frequency	IC	R-Score	P-value	Cluster
IOBC:200906087944689715	Emotional Lability	30	1.0	0.129	1.0	1
NCIT:C61486	Bulb	41	1.0	0.179	1.0	1
NCIT:C25236	Proximal	85	1.0	0.375	0.016	1
NCIT:C12742	Lower Extremity	225	1.0	1.000	0.987	2
NCIT:C25236	Proximal	139	1.0	0.616	0.987	2
NCIT:C75919	Symmetric Relationship	77	1.0	0.339	0.987	2
NCIT:C13360	Proximal	5	1.0	0.018	0.805	3
NCIT:C13360	Thoracic	5	1.0	0.018	0.805	3
IOBC:200906087944689715	Emotional Lability	73	1.0	0.321	0.002	3

3.6.2 Visualizing Semantic Descriptions

Large amounts of information become more useful when available in a comprehensive format. Visualization techniques play an important role in this regard, and provide a great deal of information in a single snap-shop of the results [7].

To make the content of semantic descriptions more understandable, the 10 annotating concepts with the highest score (or lowest, when considering P-values) in each cluster are preselected and displayed in a more convenient heatmap plot, using R's pheatmap package, which also allows row (concept) aggregation via k-means clustering. A typical example of a semantic description heatmap can be seen below in Figure 3.5.

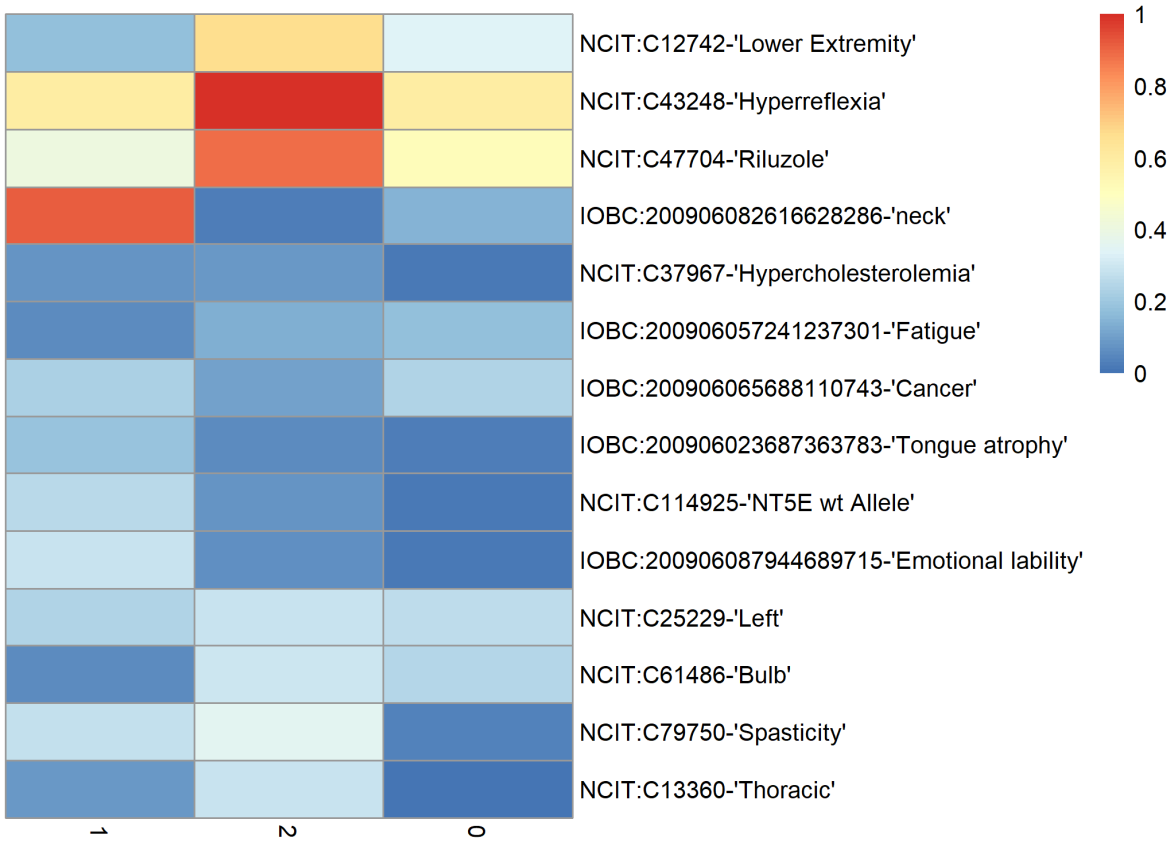

Figure 3.5: Example of a semantic description heatmap of 3 clusters, using R-scores to calculate concept relevance. Cluster ID is found below each column, and concept labels plus corresponding class IDs are on the right side of the plot. The higher the score, the greater relevancy of a concept for that cluster.

4 Evaluation

The proposed methodology was used to cluster ALS patient survey data. Its performance was assessed through three strategies: cluster quality metrics, and comparison with a non-semantic clustering approach, or against an assumed ground truth for patient groups based on the disease's progression rate.

This chapter is organized as follows. Section 4.1 introduces the patient dataset and its pre-processing. Section 4.2 explains the methods used to measure cluster quality and describes the baseline approaches. Sections 4.3 to 4.5 expound the results of ontology selection, clustering of patients, and the semantic description of data.

4.1 Case Study

ALS patient data was obtained from the LASIGE NEUROCLINOMICS2 project [90], covering both phenotypic and genotypic aspects of the disease. It comprises a dataset of 1376 ALS patients, of multiple European nationalities, who answered a standardized questionnaire upon diagnosis. The data used for this study included two sources: the questionnaire itself (questions and options for answers) and the actual answers for each patient.

4.1.1 Patient Questionnaire

A standardized questionnaire document served as the basis for all patient information. It was created in 2015 as part of the OnWebDUALS project, with the aim of collecting ALS patient data throughout European locations, to build an ALS domain ontology, and implement it in a large pan-European ALS web-database [91]. The document itself comes in a PDF format, and features 156 questions regarding patient demographic and clinical information, as well as ALS risk factors. Questions are divided by topical categories and subcategories, and are often followed by a list of optional answers the patient might choose from. Categories are expounded in greater depth in Table 4.1.

Text denoting relevant terminology was manually selected and transcribed into a list of individual terms in simple TXT format. This was done by removing stop words , punctuation, abbreviations, and detailed explanations/instructions from the content of questions, while also exercising self-discretion to recognize clinical terms. Table 4.2 gives some examples of term extraction.

Table 4.1: Content summary of questionnaire categories

Category	Title	Content
1	General Data	Identifiers of patient and recording physician, date of consultation.
2	Times	Patient age, gender, relevant dates (birth, consultation, date of 1st symptoms, diagnosis and death).
3	Disease Features	General ALS features at onset (region of first symptoms, presence of emotional lability, weight loss, cognitive symptoms, or fasciculations), handedness and predominant side of onset (in limb onset, or between UMN or LMN).
4	Clinical Signs	Specific symptoms in onset regions at study entry. Patient diagnosis is supported by revised El Escorial criteria [92], and by separate identification of Progressive muscle atrophy, Primary lateral sclerosis, Monomelic amyotrophy, and Progressive Bulbar Palsy. Neurological and emotional/depression signs are noted, with psychological diagnosis being based on Edinburgh Cognitive and Behavioural ALS Screen (ECAS) [93].
5	Disease Severity and Progression Rate	Pattern and timing of spread between regions of onset. Progression is measured using the Amyotrophic Lateral Sclerosis Functional Rating Scale (ALSFRS-R) [94], and rate of decay (FS), by finding ΔFS=48 - (current ALSFRSR/ disease duration from first symptoms to study entry).
6	Investigations	Collection of clinical tests within 6 months before study entry. This records values of Creatine Kinase, albumin, creatine, cholesterol and triglyceride measurements, medical diagnostics, and results from pulmonary function tests, including sniff nasal inspiratory pressure [95].
7	Co-Morbidities	Past and present conditions (Blood Hypertension, Metabolic, Thyroid, and Heart disease, and Cancer), and smoking habits.
8	Medication	Patient medication up to 6 months before study entry.
9	Genetic Information	Records of SOD1 and C9orf72 mutations, and familial history of ALS, FTD, Alzheimer, Parkinson, Multiple Sclerosis, (or other diseases), as well as causes of death, consanguinity and number of siblings and children.
10	Habits/Trauma and Surgery	Lifestyle choices regarding physical exercise and diet. Record of patient surgeries, blood transfusions, electric shocks, and significant emotional stress within 5 years before onset.
11	Occupations	Work occupation within 5 years before onset, demographic and environmental data on the place of living, including ALS presence. Information about the specialists who observed and diagnosed the patients, the time gap between each one another, and the tests that they ordered.

Table 4.2: Examples of term extraction from survey questions.

Question Content	Term(s)
Cognitive symptoms at onset Yes / No If yes, which (multiple choice)	"Cognitive symptoms onset"
Riluzole Date of initiation Side effects?	"Riluzole"
Mild physical exercise > 1 year (< intensity above)	"Mild physical exercise"

4.1.2 Patient dataset

The patient dataset is a CSV tabular file that compiles patient answers to the questionnaire (see Figure 4.1 for an excerpt of the data). It has 1376 rows, representing each of the patients through a unique alphanumeric code, and 631 columns, or features, based on a question or part of a question from the survey. Every entry in the dataset that contains data is identifying a patients' answer, usually relating to the presence or absence of a feature. Answers were divided into 3 types:

- **Numeric answers -** record feature values, dates, or measurements (e.g. '1', '0.5', '07/10/2015');
- **General answers -** only note the presence of features (e.g. 'yes', 'no', 'Present', 'Absent');
- **Term answers -** further describe features with semantic information, using at least a single nominal term (e.g. 'UMN', 'Upper Limb').

Entries can also be left empty, as is the case for approximately 41.5% of the dataset. Furthermore, an entry may also contain 'NR' (Not relevant), 'NA' (Not available), or 'NF' (Not feasible) to denote unreliable data, in which case they are disregarded.

ID	Dyscognition onset	Generalized onset	UMN vs LMN manifestation at onset	Limb onset	Predominant side	Predominant impairment	Fasciculations at onset	Weight loss (>10% initial weight)
	q_16f	q_16g	q_17	q_18a	q_18b	q_18c	q_19	q_20
ANTA-0001	No	No	LMN	lower limb	Left	distal	No	No
ANTA-0002	No	No	UMN	No			Yes	No
ANTA-0003	No	No	UMN	upper limb	Left	distal	No	No
ANTA-0004	No	No	LMN	upper limb	Left	distal	No	No
ANTA-0005	No	No	LMN	upper limb	Right	distal	Yes	No
ANTA-0006	No	No	LMN	upper limb	Left	distal	Yes	No
ANTA-0007	No	No	LMN	upper limb	Right	distal	No	No
ANTA-0008	No	No	LMN	lower limb	Left	distal	No	No
ANTA-0009	No	No	LMN	upper limb	Right	distal	Yes	No
ANTA-0010	No	No	LMN	lower limb	Left	distal	No	Yes
ANTA-0011	No	No	UMN	lower limb	Right	distal	Yes	No
ANTA-0012	No	No	LMN	lower limb	Right	distal	No	No
ANTA-0013	No	No	UMN	upper limb	Right	proximal	Yes	No
ANTA-0014	No	No	UMN	upper limb	Left	distal	Yes	No
ANTA-0015	No	No	UMN	lower limb	Right	distal	No	No
ANTA-0016	No	No	LMN	upper limb	Left	distal	No	No
ANTA-0017	No	No	LMN	upper limb	Symmetric	distal and proximal	Yes	No
ANTA-0018	No	No	LMN	lower limb	Left	distal	Yes	No
ANTA-0019	No	No	UMN	lower limb	Symmetric	distal	Yes	No

Figure 4.1: Excerpt of the original ALS patient dataset, questions 16 to 20. First column shows patient IDs.

Data from every patient was extracted and represented as term vector of its answers in the dataset, which converts irregular data into a standardized and more easily manageable objects. Because this work is focused on analyzing semantic data, vectors were filtered in favour of isolating, maximizing and correcting semantic information. This was done by:

1. Removing negative/neutral answers (e.g. "No", "Negative", "Normal"), blank columns, numerical data, and non-alphabetic characters;
2. Replacing positive answers (e.g. "yes", "Definite", "Positive") in data with the title of the feature, and acronyms, or initials, with full-length term labels;
3. Making a general spelling correction.

4.1.3 Challenges

Applying the proposed methodology on the patient data must deal with a number of issues, many of which are often generalized to biomedicine and EMR. These are:

1. **High Dimensionality** - The patient dataset spans several hundreds of features and distributes these unevenly across an even larger number of patients. This leads to high dimensionality problems, such as missing information and data sparseness. If patients are not being characterized by the same features, i.e. the same standards, they become more dissimilar, potentially hampering accurate clustering;
2. **Negative Data** - Over half of the dataset's entries are negative, i.e., denote the absence of a certain feature in a patient. While this kind of information can normally be mined, there are no semantic similarity or annotation methods made to handle negative semantic data;
3. **Unstructured Data** - Most of the data analysed here is textual, but is also mixed with numeric and date content as well. The survey in particular, has a freeform structure that difficults analysis, is not standardized to a data model, and comprises mostly of meaningless data. The success of the methodology will depend on well useful content was extracted from these sources;
4. **Complex and interrelated Data** - Semantic data carries intricate relationships between concepts that simpler mining approaches cannot detect. Though this work's approach was specifically designed to deal with this, the large amounts and diversity of concepts and subjects in the analysed data can still pose a challenge, if not a test to its efficacy;
5. **Interpretation of Results** - The semantic description of each patient cluster is meaningless if it cannot be contextualized within ALS symptomatology, i.e., if it makes sense to have certain concepts isolated in the same clusters. This is a difficult and subjective task since current knowledge of ALS is very limited. Furthermore, SSC is a process with many configurable parameters, and changing these can lead to different cluster results. Without a way to verify which setup is best, interpretation becomes once more uncertain.

4.2 Evaluation Approach

As stated before, cluster validation methods can measure the quality of clusters, and are used to find an optimal output of clusters and to evaluate the final results. Extrinsic validation methods rely on a ground truth i.e., reference clusters, to compare test clusters against, and assign them a score reflecting resemblance with the reference. On the other hand, intrinsic methods evaluate clusters looking only to their inherent properties, by examining how separate and compact they are [11]. Methods from both types were used throughout this chapter, via implemented algorithms in Python's Scikit-Learn module.

4.2.1 Intrinsic Validation

- **Silhouette Analysis**

  An intrinsic method that finds cluster fitness by calculating a silhouette coefficient s for all objects inside it, formally:

  $$s = \frac{b - a}{max\{b, a\}} \tag{4.1}$$

  Where b denotes object distance to other clusters, and a reflects the compactness of the cluster where the object belongs. Coefficients range from -1 to 1, where a high value suggests a good, compact and distinct cluster, and a negative one indicates an object that is in fact closer to other clusters. The quality of a cluster can be determined by the average object silhouette value [11].

  Python's Scikit-Learn also implements an optional visual representation, supporting an easier cluster assessment. A major function of silhouette plots is in helping identify the best number of clusters for a given data.

4.2.2 Extrinsic Validation

- **Fowlkes–Mallows Index**

  An extrinsic evaluation method comparing the similarities between object labels from two different clustering approaches, by calculating FM as the geometric mean between of the precision and recall of one towards the other:

  $$FM = \sqrt{\frac{TP}{TP + FP} \times \frac{TP}{TP + FN}} \tag{4.2}$$

  Where TP (True Positive) refers to the number of objects belonging to the same cluster in both approaches, while FP (False Positives) and FN (False Negatives) refer to the number of objects from one cluster being labelled into a different partition by the other approach, and vice versa [96].

- **Cluster Confusion Matrices**

  A confusion matrix (CM) is a table having, when considering a number k of clusters, a size of at least k by k. Each entry $CM_{i,j}$ indicates the number of objects of a cluster i that were labeled as being in cluster j by a different clustering approach.

  Confusion matrices are used to measure the accuracy of classifications, where a good accuracy means that most objects are represented in the diagonal of the matrix (True Positives), while the rest of the entries (the False positives and False negatives) are left nearly empty [11]. Here they are used to complement the Fowlkes–Mallows Index, by providing a way to visualize were objects labeling similarities are.

- **Evaluation through Baseline Comparison and Ground Truth**

  Comparative evaluation is necessary to understand the advantages, faults, and accuracy of SSC, and to determine if and why SSC deviates from other ground truth clusters. Two approaches were employed:

  (a) **Using a standard non-semantic clustering approach**

  Instead of using ontologies to annotate patient categorical features, and measuring semantic similarity, one-hot encoding converts these features directly into binary representations, that can be recognized and handled by clustering algorithms such as K-means. This simplistic method is meant to emulate the typical non-semantic analysis any operator could do on a patient dataset, hence its designation "Standard Approach" (SA).

  (b) **Using clinical progression groups**

  According to [97], patients can be stratified into separate groups by computing their progression rates. The progression rate is an attribute measuring how fast ALS is progressing in a patient, and is based on the recorded values in the patient's ALSFRS-R test results, through the equation:

$$ProgressionRate = \frac{48 - ALSFRSR_{1^{st}Visit}}{\Delta t_{1^{st}Symptoms;1^{st}Visit}} \tag{4.3}$$

  Where 48 is the maximum score for the ALSFRS-R scale, $ALSFRS{-}R1^{st}Visit$ is the ALSFRS-R score of a given patient at diagnosis, and $\Delta t1^{st}Symptoms; 1^{st}Visit$ is the number of months between the first symptoms and the first appointment. Patients can then be divided into 3 Progression Groups (PG), with 25% of the patients with lower or higher progression rates being grouped to create the Slow and Fast progressors groups, and the remaining 50% joined as Neutral progressors [97].

  This approach does not use clustering, nor any other data mining algorithm, and has by definition a fixed number of 3 clusters. Furthermore, some patients did not have the necessary information to calculate their progression rate, and stratification was done using 1336 patients instead. Despite this, this approach found apparent success in grouping a similar ALS patient dataset as the one analyzed here, and it was taken here as the best possible ground truth for SSC evaluation.

4.3 Results

The methodology for SSC was applied on the ALS patient data following its pre-processing stage. Outputted clusters were then examined with intrinsic and extrinsic evaluation approaches, aiming to draw an understanding on both the performance of SSC, and how ALS patients can be assorted into clinical groups.

4.3.1 Selecting Ontologies for Annotation

The process detailed in Section 3.2 was adapted here to select a list of possible ontologies to annotate the patient dataset. Survey data was used as an annotation target to test ontology performance, since it already categorized a plethora of relevant terms by category and question.

First, an initial group of 13 candidate ontologies was hand-picked from Bioportal by searching for ontologies within the "Neurological Disease" and "Neurological disorder" categories. Each of these then had their annotation potential measured against 219 terms extracted from questions and answers in the survey. Figure 4.2 (left) shows the results of single ontology annotation.

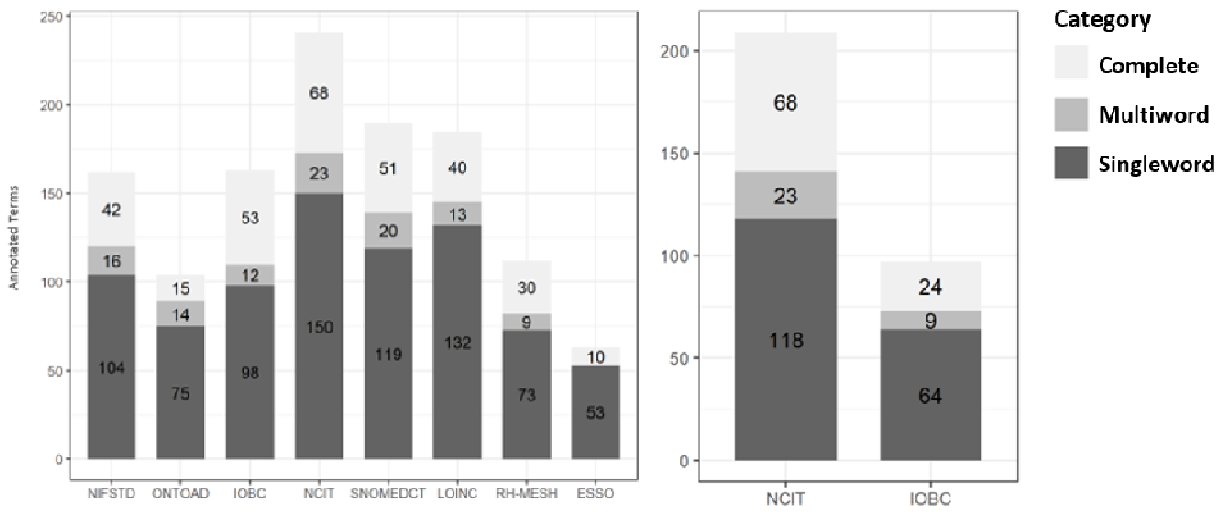

Figure 4.2: Results of ontology selection. Individual ontologies are used to annotate questionnaire terms (left). Ontologies are also tested with other top-scoring candidates to find the best annotating pair (right).

Of the original group, a total of 8 candidate ontologies were able to match the term set with a least one complete annotation. Among these, there was a general tendency correlating annotation completeness with the number of terms a given ontology could annotate. When considering both of these metrics, its also clear that the National Cancer Institute Thesaurus (NCIT) was the best-performing ontology, covering 68 terms (31% of the questionnaire term set) with complete annotations. Following NCIT, the International Ontology for Biological Components (IOBC), and the SNOMEDCT ontology each provided complete annotations for 53 and 51 terms.

A more in-depth analysis of the lowest ranking ontologies found them to be disease-specific domains (e.g. ONTOAD for Alzheimer's, and ESSO for epilepsy), which limits the scope of their content and impacts their annotation output. It seems therefore, that the heavy symptomatology component of the questionnaire is best captured by ontologies dealing in broader biomedical domains, as this latter kind consistently occupied the highest ranking positions, even in the tests involving multiple ontologies.

Next, the same set of terms was annotated with different pairs of ontologies, in order to ascertain which of these could maximize annotation coverage, and noting how well each complemented the others missing content. Multiple Ontology evaluation spanned 28 tests, all ranked in Table 4.3. The best scoring pair is shown again in Figure 4.2 (right).

In this case, results show that using both the NCIT and IOBC ontologies, 42% of the set (92 terms) was covered, out-performing all other ontology combinations. Alternatively, the joint use of NCIT-SNOMEDCT, and NCIT-NIFSTD found a favorable outcome, each respectively providing complete coverage for 86 and 75 terms. Further tests to use a triple of ontologies as an annotation source did not render a high enough coverage to justify the computational expense, and were disregarded.

Table 4.3: Ranked list of unique ontology pairs, and the total number of complete annotations they can provide for survey terms (sorted from largest to lowest).

Ontology 1	Ontology 2	Total Complete Annotations
NCIT	IOBC	92
SNOMEDCT	NCIT	86
IOBC	NIFSTD	79
RH-MESH	NCIT	79
LOINC	NCIT	77
NCIT	NIFSTD	75
LOINC	SNOMEDCT	73
SNOMEDCT	NIFSTD	72
SNOMEDCT	IOBC	72
NCIT	ONTOAD	71
LOINC	IOBC	71
ESSO	NCIT	70
LOINC	NIFSTD	61
RH-MESH	NIFSTD	61
SNOMEDCT	ONTOAD	59
RH-MESH	IOBC	59
ESSO	IOBC	59
RH-MESH	SNOMEDCT	59
IOBC	ONTOAD	58
ESSO	SNOMEDCT	55
RH-MESH	LOINC	53
ONTOAD	NIFSTD	49
LOINC	ONTOAD	48
ESSO	NIFSTD	46
ESSO	LOINC	43
RH-MESH	ONTOAD	38
ESSO	RH-MESH	38
ESSO	ONTOAD	25

4.3.2 Semantic Similarity Clustering

A unique set of 2092 terms taken from the patient dataset was annotated with the NCIT and IOBC ontologies. Annotations were sorted by their completeness into two lists of complete or single and multiword annotations. Each list was used by the semantic annotation method to convert patient term vectors into annotation vectors. Since using the FM match type only provided patients with an average of 6 annotations each, it was chosen to complement it with the CM type, increasing the average to 20 annotations and outputting a fuller semantic representation. SML computed semantic similarity between patients and the similarity scores were used to group patients into clusters, which were later semantically described.

Fine-tuning Semantic Similarity Clustering

Before the extrinsic evaluation of SSC, its overall process was optimized to output the best possible patient clusters, based on intrinsic evaluation. This meant finding which steps and variables exerted the most influence on cluster quality, and how changing them affected the final results. Three main factors can easily be singled-out by their role contribution in the methodology:

1. The completeness of patient annotations;
2. The semantic similarity measures used to compute similarity scores;
3. The clustering algorithm applied on patients and their scores.

By interpreting these factors as tunable clustering parameters, a series of tests can iteratively change their values, run SSC, observe how the patient clusters reflected the newer setup, and determine the configuration that maximizes their quality.

Cluster quality was measured primarily through silhouette analysis, as it provides the most direct way to access how cluster membership evolves with each test and what the ideal number of clusters is, without involving baseline data that would only needlessly add another layer of complexity to the evaluation. It is also the fastest method, considering it is produced as a byproduct of patient clustering algorithms. Second, by comparing cluster labels from different setups with confusion matrices and FM scores to obtain more objective and concrete numerical data.

Annotation Completeness

The completeness of annotations is one of the lowest-level parameters within SSC that can be modified, and consequently one of the most significant, given that patient semantic representation depends entirely on this factor. Two lists, of 600 complete and 4045 partial annotations were compiled from dataset annotation, and were used to derive different patient clusters with SSC. Figure 4.3 shows the silhouette plots of the clusters obtained using both approaches, and a comparison of patient labels at the same k using confusion matrices.

It can be seen that the average silhouette coefficient of clusters was small, less than 0.2, regardless of the approach used, though slightly higher using complete annotations. Clusters using partial annotations also had a noticeable portion of their patients with negative coefficients, evidencing a generally worse cluster quality.

The FMI scores were highest at $k = 2$, with a value of 0.59, but this is offset by the fact that a smaller number of clusters also increases the odds of patients being labeled into the same partition by different approaches. Other FMI scores seem to confirm this bias by drastically dropping as k increases, and confusion matrices do not show a correlation between both approaches.

Similarity Measures

Similarity measures determine how patient similarity is calculated and influence cluster membership. BMA and simGIC have been considered the best groupwise measures [30, 34], and were used here to compare patient annotation vectors. In turn, these were based on either Sanchez or Resnik IC measures (BMA used the MICA pairwise measure). Resnik's IC will give more importance to concepts less frequently found among patients, whereas the Sanchez IC employs ontology properties to deduce their semantic relevance. Both are different takes on the same question, and can therefore result in different patient similarities. Hence, four patient similarity measures were defined and evaluated:

- BMA w/IC Resnik
- BMA w/IC Sanchez
- simGIC w/IC Resnik
- simGIC w/IC Sanchez

Figure 4.4 shows the silhouette plots obtained using each measure. Once again, average silhouette coefficient scores tend to be low, and some patients have negative coefficients using BMA/IC Sanchez. Otherwise there seems to be no significant difference between either measure, even if based of different methods.

Clustering Algorithm

SSC can cluster patients using either K-means or Spectral Clustering. Figure 4.5 depicts the silhouette plot results of both algorithms, where it can be seen that Spectral Clustering performed worse than K-means, since all clusters have patients with negative coefficient values in every instance, and increasing with of k. K-means on the other hand, found a more uniform distribution of patients over clusters, which may be more adequate to solving the problem of clustering patients, despite a still low average silhouette score.

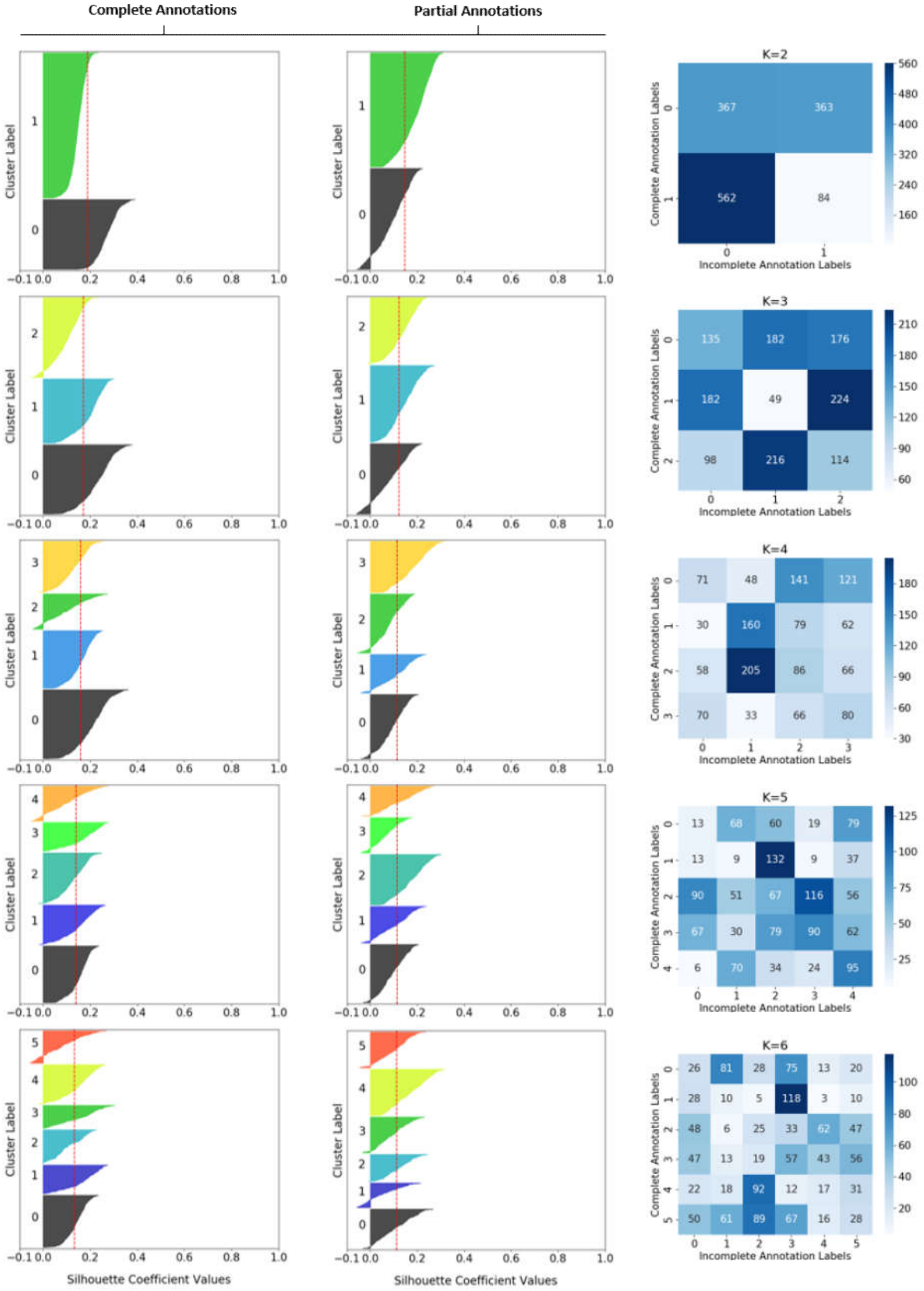

Figure 4.3: Comparing SSC clusters by the type of annotations - Patient cluster silhouettes used complete (left) or compound/partial (middle) annotations to compute patient similarity. Cluster labels from either approach are then compared via confusion matrices for every number of clusters (right). All cases used the BMA w/IC Resnik similarity measure and k-means clustering.

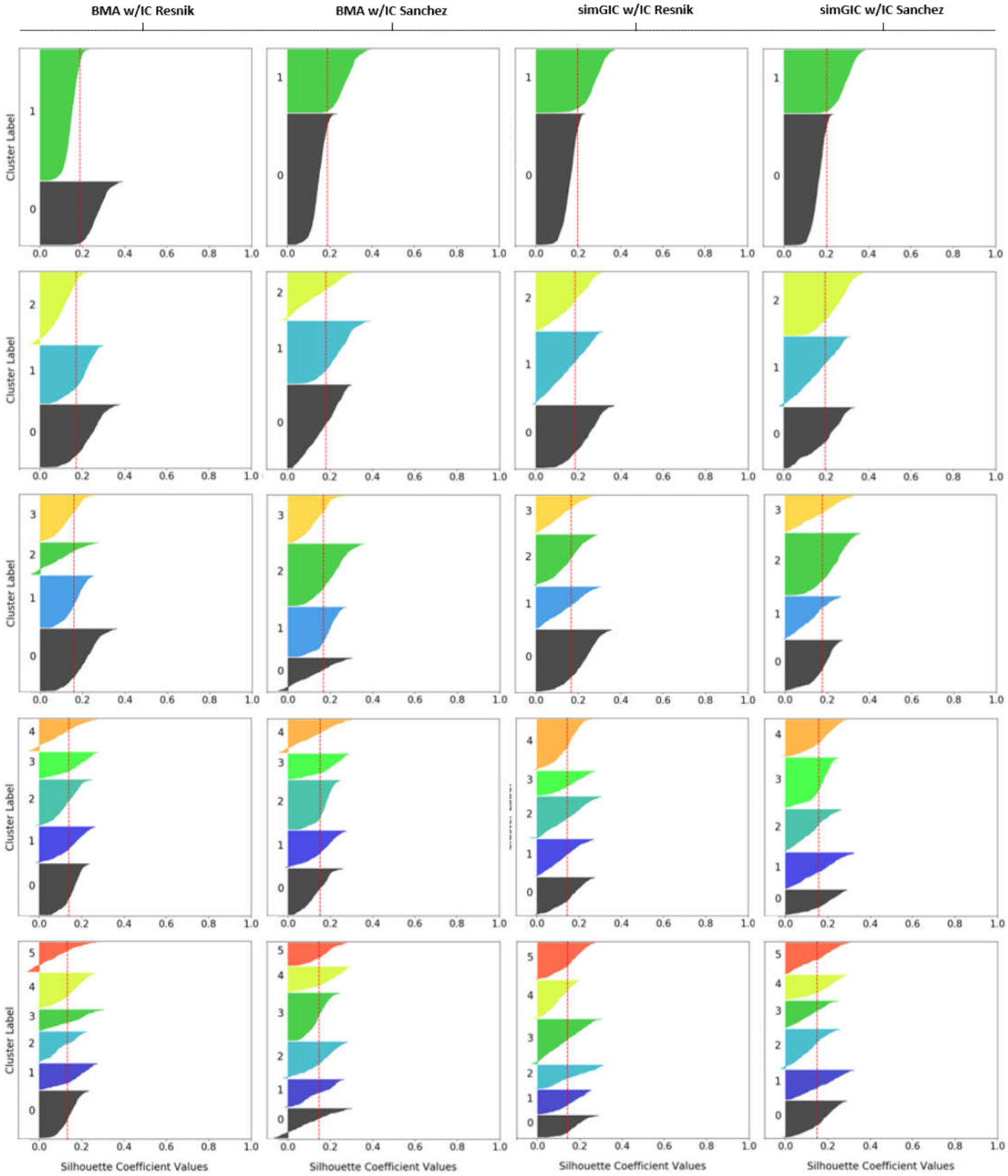

Figure 4.4: Comparing SSC clusters by score - Patient cluster silhouettes used different semantic similarity measures, from left to right: BMA w/IC Resnik; BMA w/IC Sanchez; simGIC w/IC Resnik; simGIC w/IC Sanchez. All cases used complete patient annotations and k-means clustering.

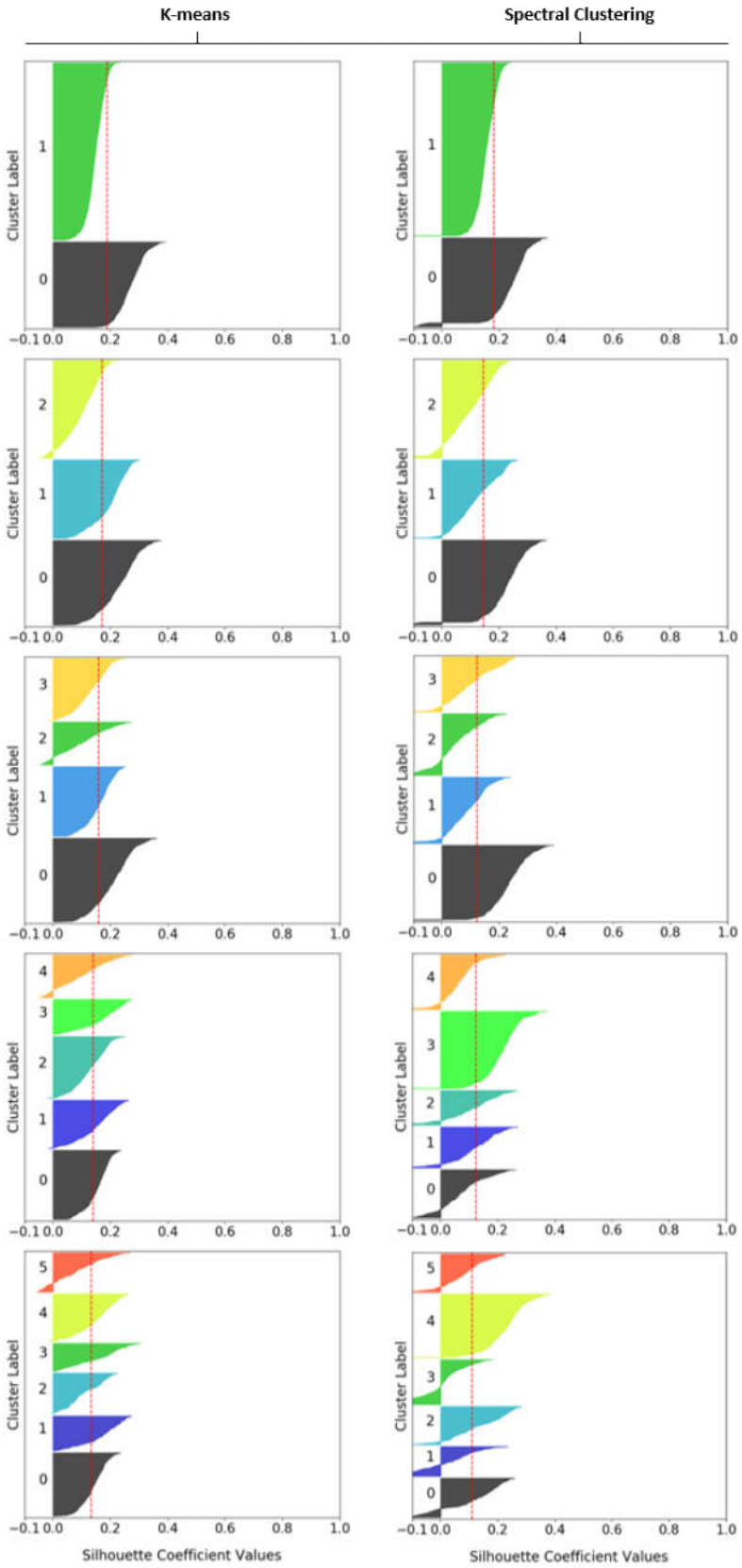

Figure 4.5: Comparing SSC clusters by clustering algorithm - Patient cluster silhouettes were generated using either the k-means (left) or spectral clustering (right) algorithm. All cases used the BMA w/IC Resnik similarity measure on complete patient annotations.

Comparison with Baselines and Ground Truth

Using the SSC method, patients were clustered using complete annotations, the BMA w/IC Resnik similarity measure, and the K-means algorithm. The resulting patient cluster labels were compared with their counterparts in either the PG or SA baseline approach, by once more computing their FMI scores and the corresponding cluster confusion matrices.

Figure 4.6 depicts the analysis results between SSC and SA labels, covering $k = \{2, ..., 6\}$. At $k = 2$, both SSC and SA grouped most patients into one cluster (69.3% in SA and 67.5% in SSC). The highest FMI score (0.58) can also be found at this stage, but as seen in previous tests, this may not be statistically significant, as label similarity between both approaches is not consistent across other cluster instances, and is in inverse proportion to k.

Generally, as k increased, SA was more conservative than SSC when reassigning patients to new clusters, and showed greater cohesiveness for specific clusters, while leaving others sparsely populated by patients ($k \geq 3$). In particular, one cluster was prevalent in all instances for including a relative majority of patients. On the other hand, SSC presented more uniform-like distributions for patients, keeping standard deviation between cluster population between 11% and 25% from k=3 to k=6, whereas the smallest deviation in SA cluster population was 43%. Because of these differences, highlighted areas in the cluster confusion matrices do more to denote the larger clusters of SA over the flat distribution in SSC, than pointing out actual similarities in patient labeling. This is once again supported by the low FMI scores.

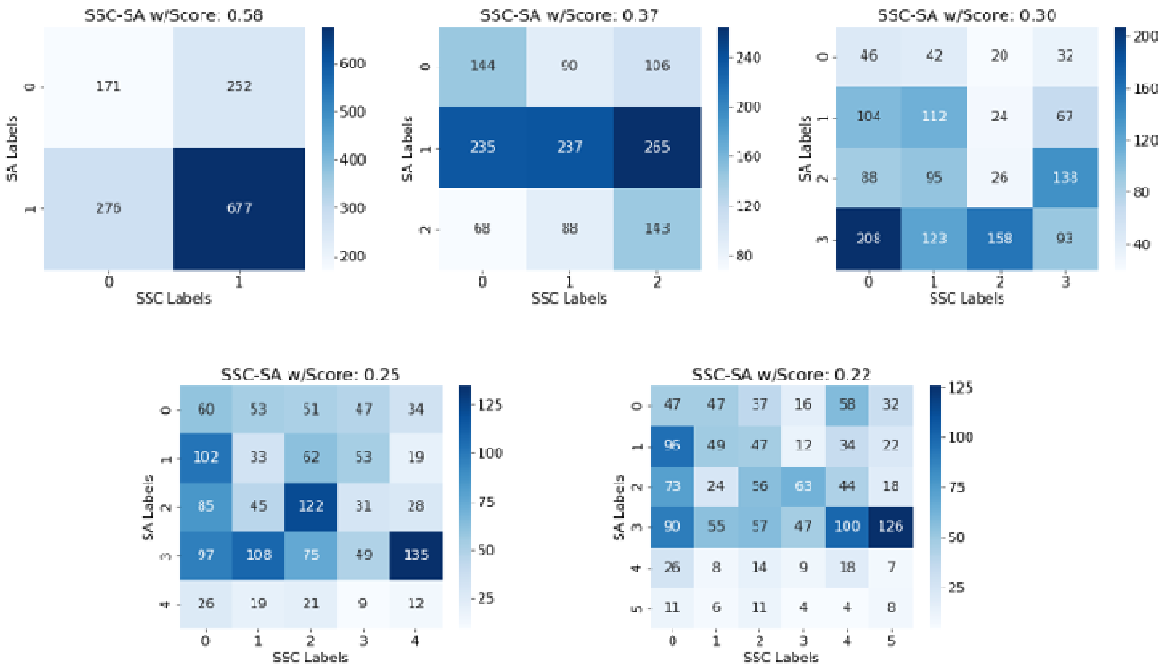

Figure 4.6: Comparing patient cluster labels between SSC and SA, by generating cluster confusion matrices for every number of clusters.

In Figure 4.7 (left), patient labels from SSC are compared against those in PG, which by definition is constricted to 3 progression groups:

- Slow Progressors (cluster 0);
- Neutral Progressors (cluster 1);
- Fast Progressors (cluster 2).

Some patients did not have enough information to calculate their progression rate, so this test proceeded with 1336 patients.

Despite this, SSC maintained the same uniform distribution as before, with a small deviation of 8.4%, while the PG cluster labels have the expected 1-2-1 ratio distribution - where 50% of patients make the Neutral progression group, and capture the largest label assignment similarity with all SSC clusters. This difference helps to explain the relatively low FMI score of 0.36, which nearly coincides with the SSC to SA counterpart in the previous analysis at k=3.

Interestingly, the SA baseline grouped nearly half (53%) of patients together into a cluster, showing apparent similarity with the ground truth PG. A closer inspection in Figure 4.7 (right) confirms that there is in fact a likeness in cluster sizes between PG and SA, but it also shows that the clusters themselves were mostly comprised of different patients, justifying the FMI score of 0.38.

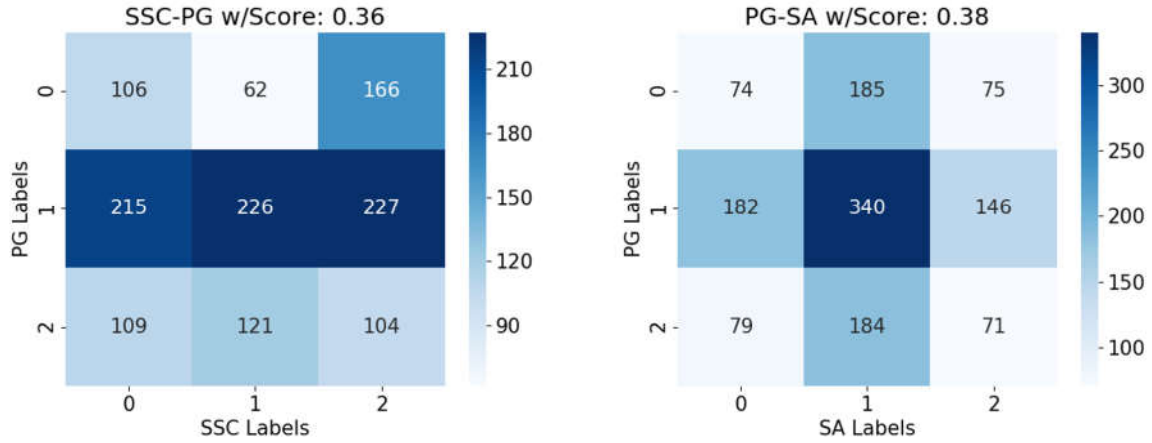

Figure 4.7: Comparison of patient cluster labels between SSC and PG (left), and between SA and PG (right).

Overall, changing clustering parameters for SSC did not significantly influence cluster outcome, nor did it improve cluster resemblance to its counterparts in PG or SA. Nonetheless, given that SA also failed to detect PG clusters, its possible that questionnaire based data is not enough to mirror patient assignments according to their progression rates.

4.3.3 Semantic Description of Clusters

Generated patient clusters were semantically described using either R-scores or P-values to measure concept relevance in clusters. Because partial annotations were frequently too generic to allow for accurate interpretation of each cluster in just 10 concepts, descriptions proceeded using only complete annotations. All SSC clusters were produced using K-means with the BMA w/IC Resnik similarity measure.

Figure 4.8 shows a heatmap of three SSC-based clusters of patient data, where R-scores were used to rank and evaluate the relevance of concepts with an $IC \geq 0.9$. The term 'Neck' was the highest scoring concept and was specific to cluster 2. This concept was traced back to the term it originality annotated, which is found on the patient dataset, under the feature "Region of onset". Thus, it could be inferred that patients from cluster 2 had an onset of ALS in the neck region, whereas others did not. The concepts 'Emotional Lability', 'Tongue Atrophy', and 'NT5E wt Allele' draw similar insights from symptoms specific to patients.

However, concepts like 'Hyporreflexia' and 'Riluzole' were found with high R-scores in all clusters. While this implicitly makes them inadequate factors to distinguish patients, they were nonetheless selected and highlighted in the plot. This is a flaw in the R-score formulation, and shows that that basing concept relevance directly on cluster frequency by itself is not an effective approach if a concept is predominant over most patients. For example, Riluzole is a common medication to treat ALS and was used in 61.5% of patients.

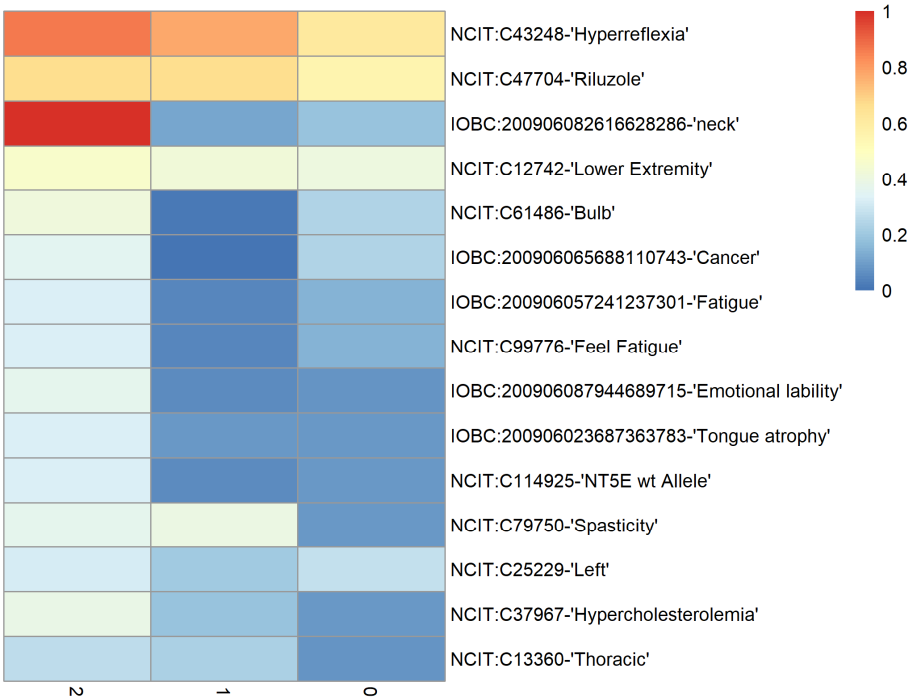

Figure 4.8: Semantic Description of generated SSC clusters using R-scores - the description relied on complete annotations from all available terms in the patient dataset.

On the other hand, Figure 4.9 shows the same SSC cluster description, but relying on P-values to rank concept relevance. Here, a more cluster-specific pattern can be seen in the general map. Concepts with greater significance (P-value ≤ 0.05), were found mostly in cluster 2. Among these, "Thoracic", "Head", "Bulb", and "Neck" are traceable to ALS onset regions, while "Fatigue","Feel Fatigue","Spasticity", "Tongue Atrophy", and "Hypercholestrolemia" relate to symptoms and co-morbidities. The concept "Mediterranean" was traced to patient diet. Cluster 0 has a highlighted, but much less significant P-value (≈ 0.3) for "Left" (side of onset) and three other concepts related to colon and prostate cancer in the patient's family. Cluster 1 had no significant concepts. Attempting to decrease the number of clusters did not improve the results for either clusters 1 or 0, and increasing it left more clusters with likewise vague and unreliable descriptions.

SSC cluster descriptions were compared with those of patient progression groups. Figure 4.10 shows the semantic description heatmap of PG clusters. Similar to cluster 2 of the previous SSC description, the Fast Progressors have several annotating concepts exclusive to them. Apart from symptoms, patients here featured "Parkinson's disease", and the concepts 'Stomach' and 'Gastric', which were traced to patient familial cancer history. The concept "neck" appears as a shared trait between Fast and Neutral Progressors, rather than unique to a single cluster. Neutral Progressors were especially prone to have 'Tongue Atrophy'. Slow Progressors were best described by the concepts 'Symmetric Relationship' and 'Left', although with a P-value ≈ 0.2.

Interestingly, SSC's cluster 2 is akin to the Fast Progressors cluster in the sense that most highlighted concepts in both heatmaps were focused on these groups. In either case, patients were described with a much larger number of ALS symptoms (e.g., 'Spasticity', 'Fatigue', 'Feel Fatigue'), and regions of onset (e.g., 'Bulb', 'Neck'). Conversely, Slow Progressors contain another subset of concepts, highlighted with much less intensity, and Neutral Progressors contain elements from the other clusters, presumably because of its larger population.

In all heatmaps, most of the relevant concepts were consistently tracked to questions from the categories 3 and 4 of the patient survey. To further explore this tendency, PG clusters were semantically described using only concepts which annotated terms from these categories.

As seen in Figure 4.11, these descriptions were more successful at finding information specific for each cluster from a limited number of concepts. Clusters included the same concepts as the previous heatmaps, but new additions can now be highlighted. In category 4, Fast Progressors were described with "Jaw clonus" and "Depression" (Pvalue ≤ 0.2), and Slow Progressors with "Hyporeflexia", which offers a curious contrast to the "Hyperreflexia" observed in Fast Progressor patients. In category 3, regions of onset are also clearly different between fast and slow progression patients, but "Neck" was no longer considered a relevant concept to either cluster.

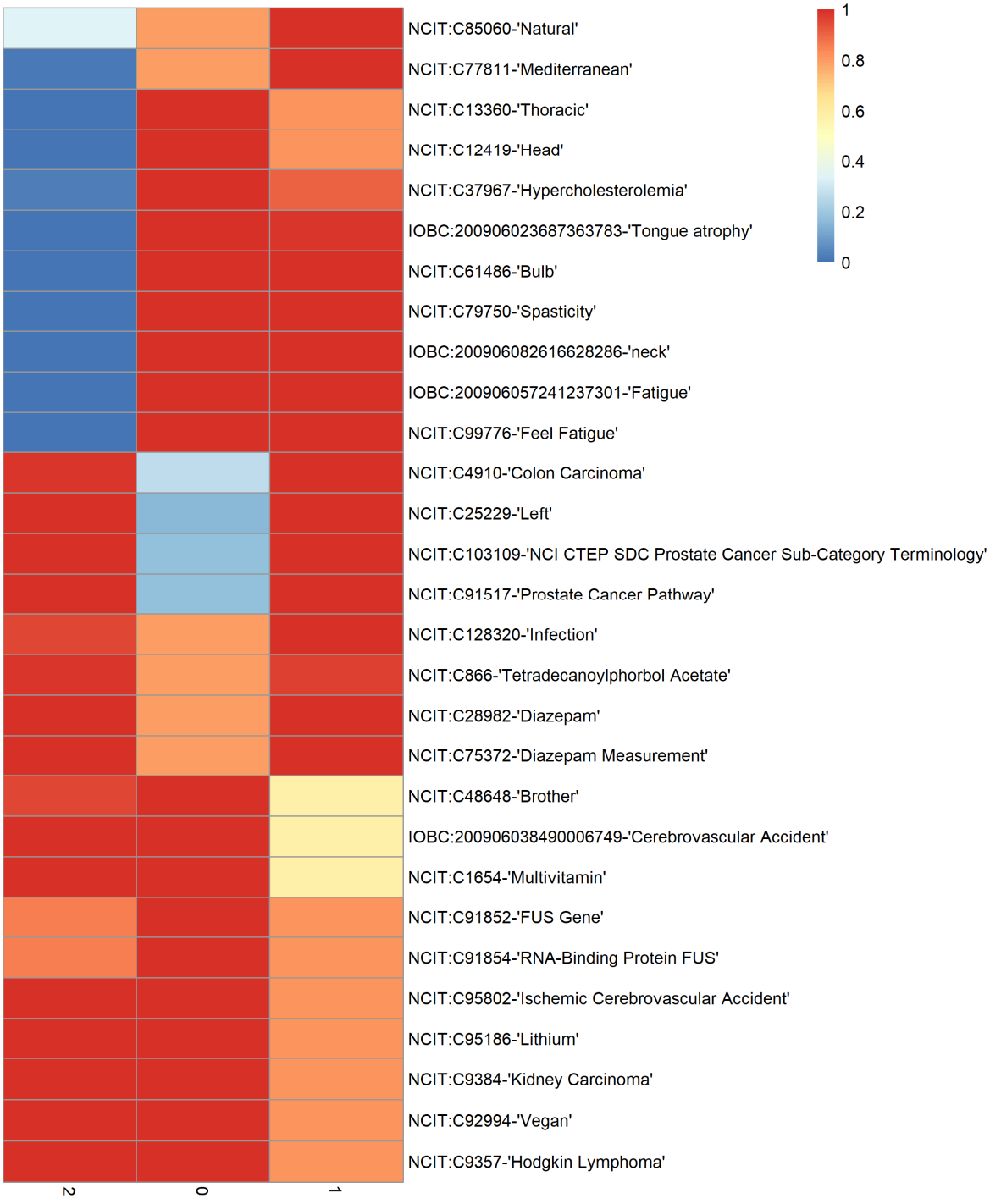

Figure 4.9: Semantic Description of generated SSC clusters using P-values, and relying on complete annotations from all terms in the patient dataset.

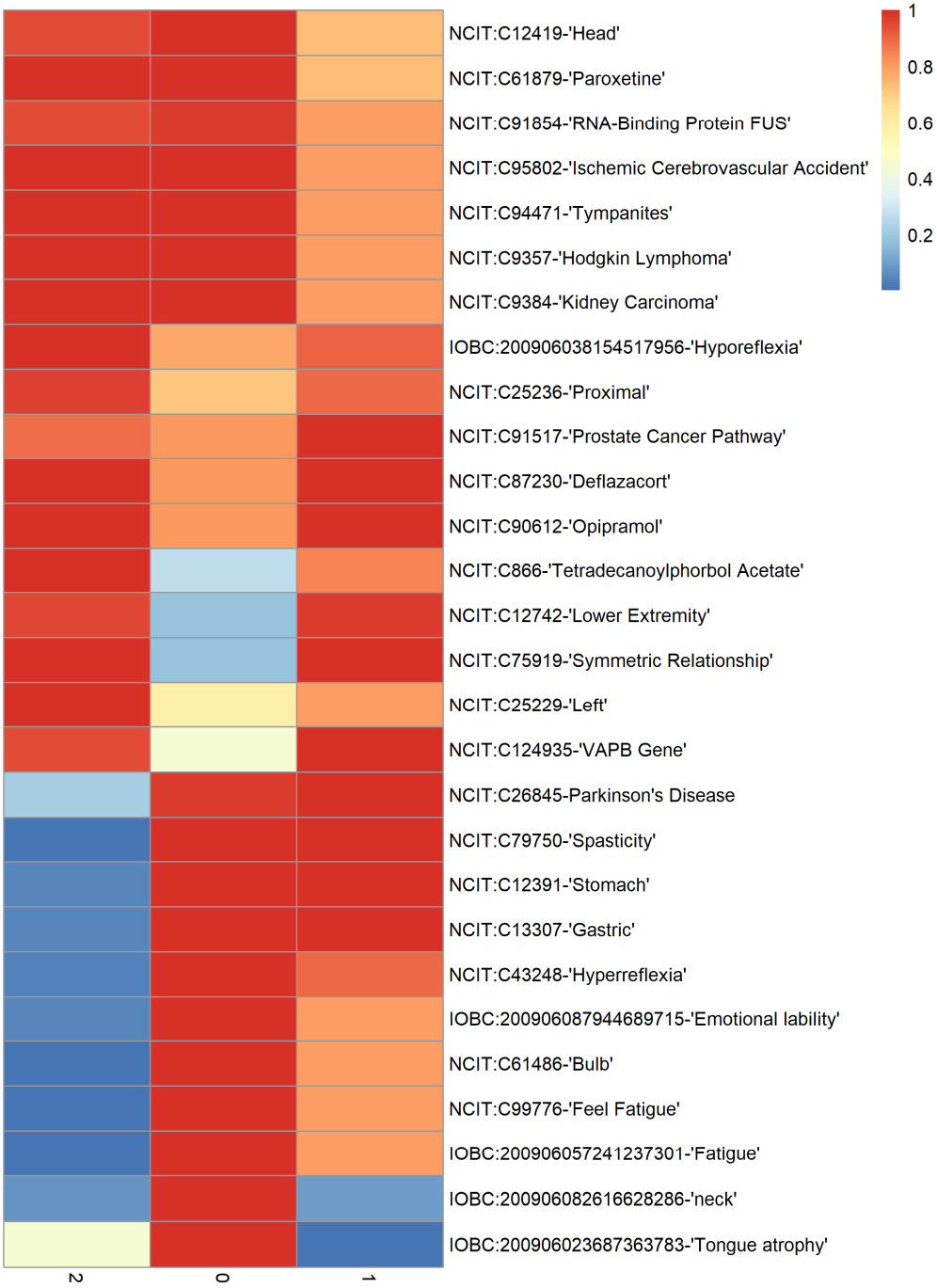

Figure 4.10: Semantic Description of ALS progression groups using P-values, and relying on complete annotations of all terms in the patient dataset.

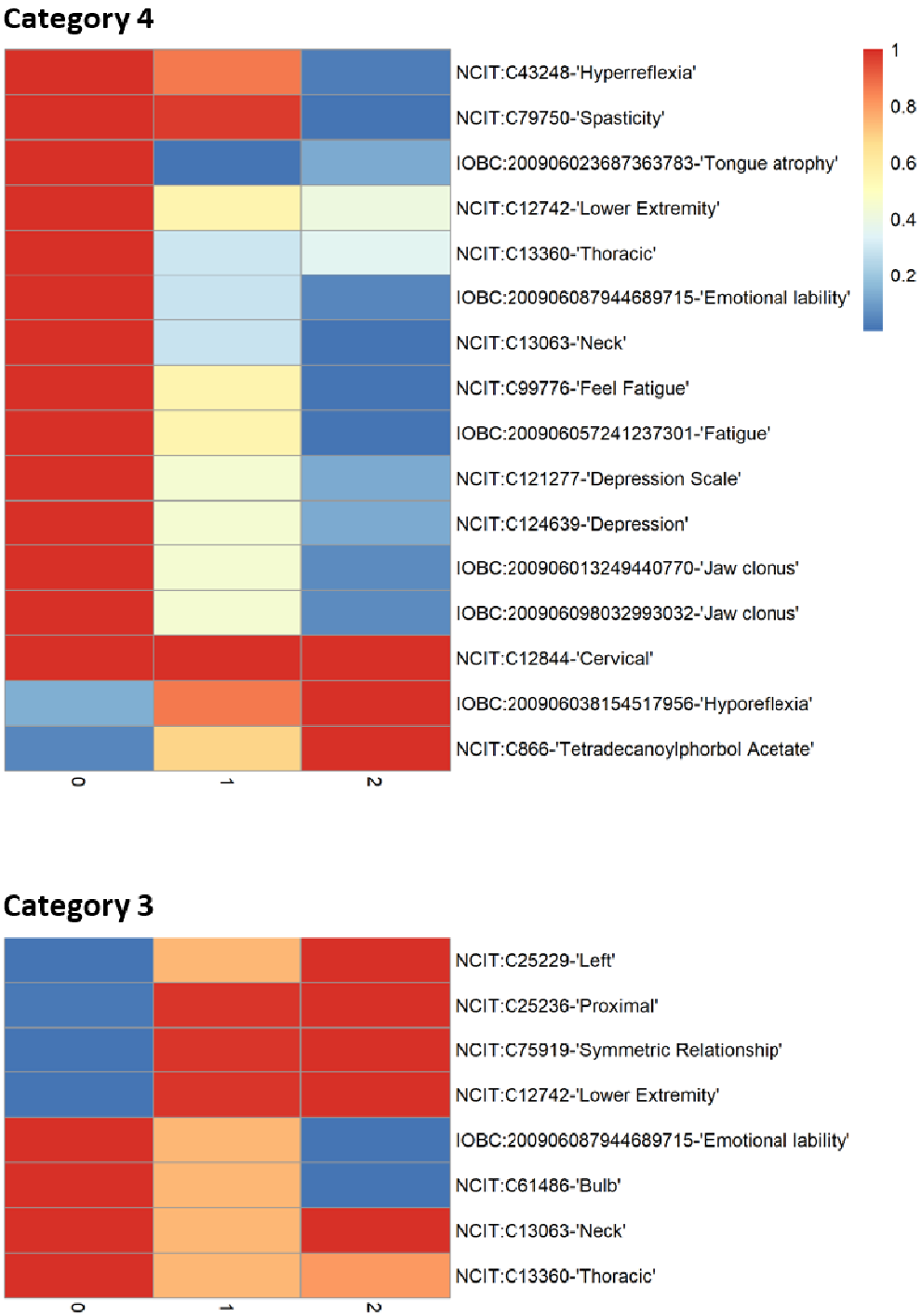

Figure 4.11: Semantic Description of ALS progression groups using P-values, relying on complete annotations from the terms in the patient dataset which referred to categories 3 (lower) and 4 (upper) of the survey.

4.4 Discussion

Cluster analysis on biomedical data with complex and unstructured features is a recognized challenge, which was here tackled through the integration of several semantic technologies and resources into a novel SSC pipeline and cluster description approach. Unlike other data mining methods, this combined approach is receptive to the semantic context hidden in text, and is robust to missing data - though it does not distinguish between missing and negative data (e.g. A non-smoker is never recognized as such). This work tested these methods on ALS patient data, dealt with the complexity of SSC by evaluating how each step and variable affected the quality of generated patient clusters, and compared the results with baselines taken over the same data to draw final conclusions.

However, the largest issue faced in using SSC was obtaining an accurate and extensive enough semantic representation of each ALS patient (i.e., many and complete annotations), not just because it found the greatest number of constrictions, but also since these were the foundation for the rest of the process. The combination of NCIT and IOBC was found to be the most suitable to annotate patient data. However, even this optimal setting only managed to cover 42% of the questionnaire terms with complete annotations. Moreover, although a questionnaire based ontology selection standardizes this procedure for other possible SSC applications on similar survey-based data, there were many terms exclusive to the patient dataset which were disregarded in this critical step (e.g., specific medicine, occupations, habits). This in turn may have indirectly contributed to a low average annotation count observed among patients. It was considered that a less accurate semantic representation of patients is preferable to a lackluster one, so the "PM" matcher type was employed to increase the patient annotation recall to an acceptable level.

Cluster validation methods demonstrated that using partial or complete annotations for data, changing semantic similarity measures, or even the number of clusters had surprisingly little influence on cluster cohesiveness and separation, and that in all cases there was a low average silhouette score (≤ 0.2). This was likely due to inaccurate or incomplete semantic representations of patients, which compromised SSC cluster quality. Alternatively, it could also mean that patient data simply did not have a natural clustering to begin with. A comparison between the patient cluster labels from SSC, SA, and PG ultimately found very little resemblance among all three, so that no definite conclusions can be made on the actual effectiveness of SSC. However, it can at least be asserted that patient feature data cannot yield the same clusters as its progression rates do.

The semantic description of clusters was able to extract better insights using P-values to rank concept relevance, than with R-scores, which were more indicative of concepts predominant over all patients. In a contrast to previous results, there was similarity among relevant concepts between SSC and the fast progression group, and in retrospect, the fact that fast progressors had a much larger subset of specific symptoms (commensurate with their stated nature) was probably what made its detection easier in the first place, which explains why SSC did not found the same relevant concepts present in slow and neutral groups. Nonetheless, this demonstrates that SSC could capture a portion of an underlying structure of ALS patient data.

It makes sense that most highlighted concepts in PG reflected the categories where there was more available data for annotations. In categories 3 and 4, there was an abundance of answers in the patient dataset, and unlike the other categories, all data in the patient dataset was also present in the survey. Making a description for each category also proved to be a reliable tactic, as differences among clusters were more easily noticeable by topic.

5 Conclusions

This project developed and evaluated a methodology to cluster patient medical records, by systematically annotating feature content with semantic meta-data from ontologies, measuring semantic similarity among patients via integration with SML, and grouping them through standard clustering strategies. This process addresses the inability of typical data mining techniques to deal with the well-known underlying complexities of unstructured biomedical data. It has also implemented an algorithm to elaborate a summary description of a cluster's semantic content while highlighting its most relevant elements. The added support for multiple ontologies also plays an important contribution to expand data integration, and in keeping the SSC on-par with the most recent developments in semantic similarity research.

A survey-based dataset of ALS patients was used as a testbed to evaluate the effectiveness of the proposed methods. It was shown that the quality of SSC clusters was mainly poor possibly due to a limited number and accuracy of annotations. This was explained by a narrow ontology coverage of concepts, and by the fact that several concepts in the dataset, but not on the survey, were not annotated. That SSC also did not acknowledge negative data also potentially contributed to a general lack of information for patient clustering. By comparison with a ground truth however, it was shown that SSC was able to identify some of the most important concepts describing clusters. Because non-semantic approaches did not come significantly closer than SSC to the expected result, the success of SSC remains an open matter for further study. Conversely, a semantic description of clusters grounded on enrichment analysis was able to consistently provide meaningful insights on patient stratification. Cluster descriptions are particularly useful since they can be applied to any patient clustering, semantic or not. The description of reference clusters evidenced how symptoms and regions of onset can suggest the progression speed of ALS. The ability to quickly characterize patient groups presents the largest contributions of this work towards personalized medicine, which relies on careful and detailed diagnostics of each patient.

5.1 Future Work

The methods presented in this work represent a first approach to enable semantic similarity clustering over multi-domain and complex biomedical data. Future works can seek to improve upon its shortcomings, implement new functionalities, or expand on previous. They can focus on:

- Investigating how negative data can be modeled into a negative kind of annotations, and develop a semantic similarity measure to handle both kinds of information;
- Improving the overall implementation of the methodology by exploring and integrating other software applications. For example, the recent ViSEAGO R package [98] offers a single source compiling functions to compute semantic simialrity, cluster data through similarity scores, and simultaneously produce an enrichment analysis;

- Creating a semantic data model for the survey data, i.e., restructuring survey content into a knowledge representation schema, and describing the meaning of concepts by means of mapping them to ontology definitions. This would provide an optimized knowledge background with which to annotate the patient data;
- Further develop the scoring functions of semantic descriptions, by reformulating R-Scores, and extending P-values to include semantic context in ontologies;
- Combine numerical data into the clustering process.

Furthermore, the methodology can still be applied on other datasets with semantically annotatable biomedical features, which would open the opportunity to reevaluate the influence of metrics, annotation types, and clustering algorithms on cluster results.

Bibliography

[1] Riccardo Bellazzi, Marianna Diomidous, Indra Sarkar, Katsuhiko Takabayashi, Andreas Ziegler, and Alexa T. McCray. Data analysis and data mining: Current issues in biomedical informatics. *Methods of information in medicine*, 50:536–44, 12 2011.

[2] Jelena Jovanović and Ebrahim Bagheri. Semantic annotation in biomedicine: the current landscape. *Journal of Biomedical Semantics*, 8(1), 2017.

[3] Herbert A. Edelstein. Introduction to data mining and knowledge discovery third edition by two crows corporation introduction to data mining and knowledge discovery. 1999.

[4] Subhash Chandra Pandey. Data mining techniques for medical data: A review. *2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES)*, pages 972–982, 2016.

[5] Holger Fröhlich, Rudi Balling, Niko Beerenwinkel, Oliver Kohlbacher, Santosh Kumar, Thomas Lengauer, Marloes H. Maathuis, Yves Moreau, Susan A. Murphy, Teresa Przytycka, Michael Rebhan, Hannes Röst, Andreas Schuppert, Matthias Schwab, Rainer Spang, Daniel Stekhoven, Jimeng Sun, Andreas Weber, Daniel Ziemek, and Blaz Zupan. From hype to reality: Data science enabling personalized medicine. *BMC Medicine*, 16, 12 2018.

[6] Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T Dudley. Deep learning for healthcare: review, opportunities and challenges. *Briefings in bioinformatics*, 19, 05 2017.

[7] Krzysztof J. Cios and G. William Moore. Uniqueness of medical data mining. *Artificial intelligence in medicine*, 26 1-2:1–24, 2002.

[8] Nicholas Sioutos, Sherri de Coronado, Margaret W. Haber, Frank W. Hartel, Wen-Ling Shaiu, and Lawrence W. Wright. Nci thesaurus: A semantic model integrating cancer-related clinical and molecular information. *J. of Biomedical Informatics*, 40(1):30–43, February 2007.

[9] R. Hoehndorf, P. N. Schofield, and G. V. Gkoutos. The role of ontologies in biological and biomedical research: a functional perspective. *Briefings in Bioinformatics*, 16(6):1069–1080, 2015.

[10] Ian Horrocks. Ontologies and the semantic web. *Commun. ACM*, 51(12):58–67, December 2008.

[11] Jiawei Han, Micheline Kamber, and Jian Pei. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd edition, 2011.

[12] Catia Pesquita. *Semantic Similarity in the Gene Ontology*, pages 161–173. Springer New York, 2017.

[13] João D. Ferreira and Francisco M. Couto. Multi-domain semantic similarity in biomedical research. *BMC Bioinformatics*, 20(10):246, 2019.

[14] Vincent Grollemund, Pierre-François Pradat, Giorgia Querin, Francois Delbot, Gaétan Le Chat, Jean-François Pradat-Peyre, and Peter Bede. Machine learning in amyotrophic lateral sclerosis: Achievements, pitfalls, and future directions. *Frontiers in Neuroscience*, 13, 02 2019.

[15] Stephen R. Pfohl, Renaid B. Kim, Grant S. Coan, and Cassie S. Mitchell. Unraveling the complexity of amyotrophic lateral sclerosis survival prediction. In *Front. Neuroinform.*, 2018.

[16] Jingshan Huang, Dejing Dou, Jiangbo Dang, J Harold Pardue, Xiao Qin, Jun Huan, William T Gerthoffer, and Ming Tan. Knowledge acquisition, semantic text mining, and security risks in health and biomedical informatics, Feb 2012.

[17] Thabet Slimani. Semantic annotation: The mainstay of semantic web. *International Journal of Computer Applications Technology and Research*, 2(6):763–770, 2013.

[18] Stephan Grimm, Andreas Abecker, Johanna Völker, and Rudi Studer. Ontologies and the semantic web. In *Handbook of Semantic Web Technologies*, 2011.

[19] Gayathri R and Uma Vijayasundaram. Ontology based knowledge representation technique, domain modeling languages and planners for robotic path planning: A survey. *ICT Express*, 4, 04 2018.

[20] Lawrence E. Hunter. Knowledge-based biomedical data science. *Data Science*, 1:1–7, 03 2017.

[21] Christine Golbreich, Matthew Horridge, Ian Horrocks, Boris Motik, and Rob Shearer. *OBO and OWL: Leveraging Semantic Web Technologies for the Life Sciences*, volume 4825, pages 169–182. 01 2007.

[22] Embl-Ebi. Gene ontology and go annotations. `https://www.ebi.ac.uk/QuickGO/term/GO:0005634`.

[23] Jonathan M Mortensen, Matthew Horridge, Mark Musen, and Natasha Noy. Applications of ontology design patterns in biomedical ontologies. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, 2012:643–52, 11 2012.

[24] Daniel Faria, Catia Pesquita, Isabela Mott, Catarina Martins, Francisco Couto, and Isabel F. Cruz. Tackling the challenges of matching biomedical ontologies. *Journal of Biomedical Semantics*, 9, 12 2018.

[25] Xingsi Xue, Zhi Hang, and Zhengyi Tang. Interactive biomedical ontology matching. *PLOS ONE*, 14:e0215147, 04 2019.

[26] O Bodenreider. Biomedical ontologies in action: role in knowledge management, data integration and decision support, 2008.

[27] Alan R. Aronson and François-Michel Lang. An overview of metamap: historical perspective and recent advances. *Journal of the American Medical Informatics Association : JAMIA*, 17 3:229–36, 2010.

[28] Clement Jonquet, Nigam Shah, and Mark Musen. The open biomedical annotator. *Summit on translational bioinformatics*, 2009:56–60, 03 2009.

[29] Sébastien Harispe, Sylvie Ranwez, Stefan Janaqi, and Jacky Montmain. Semantic similarity from natural language and ontology analysis. *CoRR*, abs/1704.05295, 2017.

[30] Pietro Hiram Guzzi, Marco Mina, Concettina Guerra, and Mario Cannataro. Semantic similarity analysis of protein data: assessment with biological features and issues. *Briefings in bioinformatics*, 13 5:569–85, 2012.

[31] Philip Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1*, IJCAI'95, pages 448–453, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc.

[32] David Sánchez, Albert Solé-Ribalta, Montserrat Batet, and Francesc Serratosa. Enabling semantic similarity estimation across multiple ontologies: An evaluation in the biomedical domain. *Journal of Biomedical Informatics*, 45(1):141–155, 2012.

[33] Andreas Schlicker, Francisco S. Domingues, Jörg Rahnenführer, and Thomas Lengauer. A new measure for functional similarity of gene products based on gene ontology. *BMC Bioinformatics*, 7:302 – 302, 2006.

[34] Catia Pesquita, Daniel Faria, Hugo Bastos, António Ferreira, André Falcão, and Francisco Couto. Metrics for go based protein semantic similarity: A systematic evaluation. *BMC bioinformatics*, 9 Suppl 5:S4, 02 2008.

[35] Waqar Ali and Charlotte M. Deane. Functionally guided alignment of protein interaction networks for module detection. *Bioinformatics*, 25(23):3166–3173, 2009.

[36] Young-Rae Cho, Woochang Hwang, Murali Ramanathan, and Aidong Zhang. Semantic integration to identify overlapping functional modules in protein interaction networks. *BMC Bioinformatics*, 8(1), 2007.

[37] M. Popescu, J.m. Keller, and J.a. Mitchell. Fuzzy measures on the gene ontology for gene product similarity. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 3(3):263–274, 2006.

[38] David Martin, Christine Brun, Elisabeth Remy, Pierre Mouren, Denis Thieffry, and Bernard Jacq. Gotoolbox: Functional analysis of gene datasets based on gene ontology. *Genome biology*, 5:R101, 02 2004.

[39] Hongwei Wu, Zhengchang Su, Fenglou Mao, Victor Olman, and Ying Xu. Prediction of functional modules based on comparative genome analysis and gene ontology application, May 2005.

[40] Sidahmed Benabderrahmane, Malika Smail-Tabbone, Olivier Poch, Amedeo Napoli, and Marie-Dominique Devignes. Intelligo: a new vector-based semantic similarity measure including annotation origin, Dec 2010.

[41] Da Wei Huang, Brad T Sherman, Qina Tan, Jack R Collins, W Gregory Alvord, Jean Roayaei, Robert Stephens, Michael W Baseler, H Clifford Lane, Richard A Lempicki, and et al. The david gene functional classification tool: a novel biological module-centric algorithm to functionally analyze large gene lists, 2007.

[42] Meeta Mistry and Paul Pavlidis. Gene ontology term overlap as a measure of gene functional similarity ¡http://www.biomedcentral.com/1471-2105/9/327. *BMC bioinformatics*, 9:327, 02 2008.

[43] Hisham Al-Mubaid and Anurag Nagar. Comparison of four similarity measures based on go annotations for gene clustering. *2008 IEEE Symposium on Computers and Communications*, pages 531–536, 2008.

[44] S Falcon and R Gentleman. Using GOstats to test gene lists for GO term association. *Bioinformatics*, 23(2):257–8, 2007.

[45] Ping Ye, Brian D. Peyser, Xuewen Pan, Jef D. Boeke, Forrest A. Spencer, and Joel S. Bader. Gene function prediction from congruent synthetic lethal interactions in yeast. *Molecular Systems Biology*, 1:2005.0026 – 2005.0026, 2005.

[46] Brendan Sheehan, Aaron J. Quigley, Benoit Gaudin, and Simon A. Dobson. A relation based measure of semantic similarity for gene ontology annotations. *BMC Bioinformatics*, 9:468 – 468, 2008.

[47] Homin K. Lee, Amy Kuang-Hua Hsu, Jon Sajdak, Jie Qin, and Paul Pavlidis. Coexpression analysis of human genes across many microarray data sets. *Genome research*, 14 6:1085–94, 2004.

[48] Haiyuan Yu, Ronald Jansen, and Mark Gerstein. Developing a similarity measure in biological function space. *Bioinformatics*, online:1–18, 2007.

[49] Julie Chabalier, Jean Mosser, and Anita Burgun. A transversal approach to predict gene product networks from ontology-based similarity. *BMC bioinformatics*, 8:235, 02 2007.

[50] Olivier Bodenreider, Marc Aubry, and Anita Burgun-Parenthoine. Non-lexical approaches to identifying associative relations in the gene ontology. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 91–102, 2004.

[51] James Z Wang, Zhidian Du, Rapeeporn Payattakool, Philip Yu, and Chin-Fu Chen. A new method to measure the semantic similarity of go terms. *Bioinformatics (Oxford, England)*, 23:1274–81, 06 2007.

[52] Francisco Couto, Mário Silva, and Pedro Coutinho. Measuring semantic similarity between gene ontology terms. *Data & Knowledge Engineering*, 61:137–152, 04 2007.

[53] Jay J. Jiang and David W. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. *CoRR*, cmp-lg/9709008, 1997.

[54] Dekang Lin. An information-theoretic definition of similarity. In *Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98, pages 296–304, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc.

[55] Razib M. Othman, Safaai Deris, and Rosli Md. Illias. A genetic similarity algorithm for searching the gene ontology terms and annotating anonymous protein sequences. *Journal of biomedical informatics*, 41 1:65–81, 2008.

[56] Viktor Pekar and Steffen Staab. Taxonomy learning - factoring the structure of a taxonomy into a semantic classification decision. In *COLING 2002: The 19th International Conference on Computational Linguistics*, 2002.

[57] Xiaomei Wu, Lei Zhu, Jie Guo, Da-Yong Zhang, and Kui Lin. Prediction of yeast protein-protein interaction network: insights from the gene ontology and annotations, Apr 2006.

[58] Hui Yu, Lei Gao, Kang Tu, and Zheng Guo. Broadly predicting specific gene functions with expression similarity and taxonomy similarity. *Gene*, 352:75–81, 2005.

[59] Bo Li, James Z. Wang, Frank Alex Feltus, Jizhong Zhou, and Feng Luo. Effectively integrating information content and structural relationship to improve the go-based similarity measure between proteins. *CoRR*, abs/1001.0958, 2010.

[60] Francisco Couto, Mário Silva, and Pedro Coutinho. Implementation of a functional semantic similarity measure between gene-products. 01 2003.

[61] Shobhit Jain and Gary D Bader. An improved method for scoring protein-protein interactions using semantic similarity within the gene ontology. *BMC bioinformatics*, 11:562, 11 2010.

[62] Hongwei Wu, Zhengchang Su, Fenglou Mao, Victor Olman, and Ying Xu. Prediction of functional modules based on comparative genome analysis and gene ontology application. *Nucleic Acids Research*, 33:2822 – 2837, 2005.

[63] Zhibiao Wu and Martha Palmer. Verbs semantics and lexical selection. In *Proceedings of the 32Nd Annual Meeting on Association for Computational Linguistics*, ACL '94, pages 133–138, Stroudsburg, PA, USA, 1994. Association for Computational Linguistics.

[64] Antonio Sanfilippo, Christian Posse, Banu Gopalan, Rick Riensche, Nathaniel Beagley, Bob Baddeley, Stephen Tratz, and Michelle Gregory. Combining hierarchical and associative gene ontology relations with textual evidence in estimating gene and gene product similarity. *NanoBioscience, IEEE Transactions on*, 6:51 – 59, 04 2007.

[65] M.a. Rodriguez and M.j. Egenhofer. Determining semantic similarity among entity classes from different ontologies. *IEEE Transactions on Knowledge and Data Engineering*, 15(2):442–456, 2003.

[66] Abraham Bernstein, Esther Kaufmann, Christoph Kiefer, and Christoph Bürki. Simpack: A generic java library for similarity measures in ontologies. 08 2006.

[67] Giuseppe Pirrò. A semantic similarity metric combining features and intrinsic information content. *Data & Knowledge Engineering*, 68:1289–1308, 11 2009.

[68] Jiang Li, Binsheng Gong, Xi Chen, Tao Liu, Chao Wu, Fan Zhang, Chunquan li, Xiang li, Shaoqi Rao, and Xia Li. Dosim: An r package for similarity between diseases based on disease ontology. *BMC bioinformatics*, 12:266, 06 2011.

[69] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to Data Mining, (First Edition)*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2005.

[70] Minlei Liao, Yunfeng Li, Farid Kianifard, Engels Obi, and Stephen Arcona. Cluster analysis and its application to healthcare claims data: a study of end-stage renal disease patients who initiated hemodialysis. *BMC Nephrology*, 17, 12 2016.

[71] David Arthur and Sergei Vassilvitskii. K-means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '07, pages 1027–1035, Philadelphia, PA, USA, 2007. Society for Industrial and Applied Mathematics.

[72] Ulrike Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, December 2007.

[73] Benjamin Auffarth. Spectral Graph Clustering. pages 1–12, 2007.

[74] Jake VanderPlas. *Python Data Science Handbook: Essential Tools for Working with Data*. O'Reilly Media, Inc., 1st edition, 2016.

[75] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[76] Sebastian Raschka. Mlxtend: Providing machine learning and data science utilities and extensions to python's scientific computing stack. *The Journal of Open Source Software*, 3(24), April 2018.

[77] Jared P. Lander. *R for Everyone: Advanced Analytics and Graphics (2Nd Edition)*. Addison-Wesley Professional, 2nd edition, 2017.

[78] Rstudio, new open-source ide for r. `https://blog.rstudio.com/2011/02/28/rstudio-new-open-source-ide-for-r/`.

[79] Kristian Ovaska, Marko Laakso, and Sampsa Hautaniemi. Fast gene ontology based clustering for microarray experiments. *BioData Mining*, 1:11 – 11, 2008.

[80] Sidahmed Benabderrahmane and Hayet Mekami. Ontology-based gene set enrichment analysis using an efficient semantic similarity measure and functional clustering. 867:151–159, 01 2012.

[81] Zhiwen Yu, Wei Luo, Guangyuan Fu, and Jun Wang. Interspecies gene function prediction using semantic similarity. *BMC Systems Biology*, 10:495–507, 12 2016.

[82] Marek Ostaszewski, Emmanuel Kieffer, Grégoire Danoy, Reinhard Schneider, and Pascal Bouvry. Clustering approaches for visual knowledge exploration in molecular interaction networks. *BMC Bioinformatics*, 19, 12 2018.

[83] S. S. Desai and J. A. Laxminarayana. Wordnet and semantic similarity based approach for document clustering. In *2016 International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS)*, pages 312–317, Oct 2016.

[84] Christos Bouras and Vassilis Tsogkas. W-kmeans: Clustering news articles using wordnet. In Rossitza Setchi, Ivan Jordanov, Robert J. Howlett, and Lakhmi C. Jain, editors, *Knowledge-Based and Intelligent Information and Engineering Systems*, pages 379–388, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg.

[85] Ilya Blokh and Vassil Alexandrov. News clustering based on similarity analysis. *Procedia Computer Science*, 122:715–719, 01 2017.

[86] Sarah Westbury, Ernest Turro, Daniel Greene, Claire Lentaigne, Anne Kelly, Tadbir Bariana, Ilenia Simeoni, Xavier Pillois, Antony Attwood, Steven Austin, Sjoert Jansen, Tamam Bakchoul, Abi Crisp-Hihn, Wendy N Erber, Rémi Favier, Nicola Foad, Michael Gattens, Jennifer D Jolley, Ri Liesner, and Kathleen Freson. Human phenotype ontology annotation and cluster analysis to unravel genetic defects in 707 cases with unexplained bleeding and platelet disorders. *Genome Medicine*, 7, 04 2015.

[87] Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, Scott L. Pomeroy, Todd R. Golub, Eric S. Lander, and et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles, Oct 2005.

[88] Guangchuang Yu, Li-Gen Wang, Yanyan Han, and Qing-Yu He. clusterprofiler: an r package for comparing biological themes among gene clusters. *OMICS: A Journal of Integrative Biology*, 16(5):284–287, 2012.

[89] Daniel Faria, Catia Pesquita, Emanuel Santos, Matteo Palmonari, Isabel F. Cruz, and Francisco M. Couto. The agreementmakerlight ontology matching system. In *OTM Conferences*, 2013.

[90] Neuroclinomics2 - unravelling prognostic markers in neurodegenerative diseases through clinical and omics data integration (neuroclinomics2). *INESCID*.

[91] Maria Piotrkiewicz, Mamede de carvalho, Peter M. Andersen, Julian Grosskreutz, Magdalena Kuzma, Susanne Petri-Mals, and Teresa Podsiadły-Marczykowska. Onwebduals: the european project funded by national agencies under the patronage of joint programme – neurodegenerative disease research (jpnd). 06 2016.

[92] Albert Christian Ludolph, Vivian Drory, Orla Hardiman, Imaharu Nakano, John Ravits, Wim Robberecht, and Jeremy M Shefner. A revision of the el escorial criteria - 2015. *Amyotrophic lateral sclerosis & frontotemporal degeneration*, 16 5-6:291–2, 2015.

[93] Sharon Abrahams, Judith Newton, Elaine Niven, Jennifer A. Foley, and Thomas Bak. Screening for cognition and behaviour changes in als. *Amyotrophic lateral sclerosis & frontotemporal degeneration*, 15, 06 2013.

[94] Jesse Michael Cedarbaum, Nancy Stambler, Errol Malta, Cynthia L. Fuller, and . BDNFALSStudyGroup1AcompletelistingoftheBDNFStudyGr. The alsfrs-r: a revised als functional rating scale that incorporates assessments of respiratory function. *Journal of the Neurological Sciences*, 169:13–21, 1999.

[95] Marta Kaminska, Francine Noel, and Basil Petrof. Optimal method for assessment of respiratory muscle strength in neuromuscular disorders using sniff nasal inspiratory pressure (snip). *PLOS ONE*, 12:e0177723, 05 2017.

[96] Maria Halkidi, Yannis Batistakis, and Michalis Vazirgiannis. On clustering validation techniques. *Journal of Intelligent Information Systems*, 17, 10 2001.

[97] Sofia Pires, Marta Gromicho, Susana Pinto, Mamede de carvalho, and Sara C . Madeira. Predicting non-invasive ventilation in als patients using stratified disease progression groups. pages 748–757, 11 2018.

[98] Aurélien Brionne, Amélie Juanchich, and Christelle Hennequet-Antier. Viseago: a bioconductor package for clustering biological functions using gene ontology and semantic similarity. *BioData Mining*, 12, 12 2019.